

Statistical Basis of Biomedical Investigation

John Sorkin M.D., Ph.D.

Chief, Biostatistics and Informatics

Baltimore VA GRECC

And

University of Maryland Claude Pepper
Older Americans Independence Center

Dedicated To

- Dr. Reubin Andres
 - Great man of science
 - True gentleman
 - My mentor

References

- PDQ Statistics 3rd edition, Norman G, Streiner D, BC Decker Inc, Hamilton, 2003
 - Short, quick, sweet, easy to understand
 - <https://www.yumpu.com/en/document/view/5047707/pdq-statistics-third-edition-faculty-of-health-sciences-mcmaster->
- Using and Understanding Medical Statistics 5th edition, Matthews DE, Farewell VT, Krager, Basel, 2015
 - Slightly longer, easy to understand, more “formal”
 - <https://www.karger.com/Article/Pdf/99416>

References

- Biostatistics: A Foundation for Analysis in the Health Sciences 10th edition, Daniel WW, Cross L, John Wiley & Sons 2013
 - Great textbook, comprehensive, easy to understand (for a textbook)

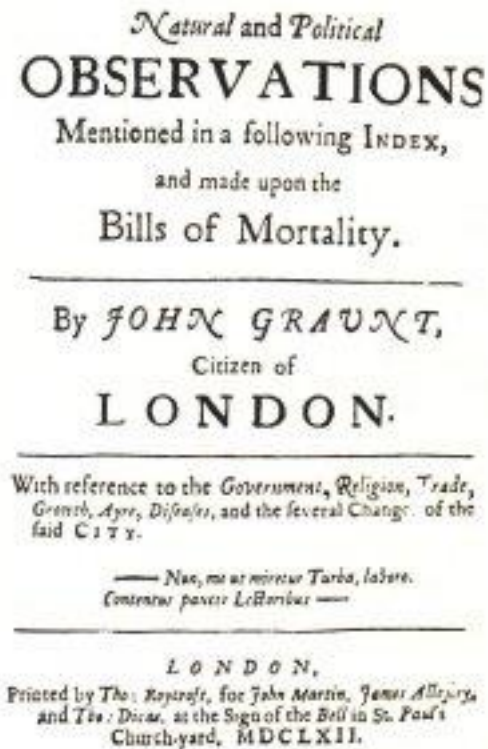
A little History

Pioneers

John Graunt

1620-1674

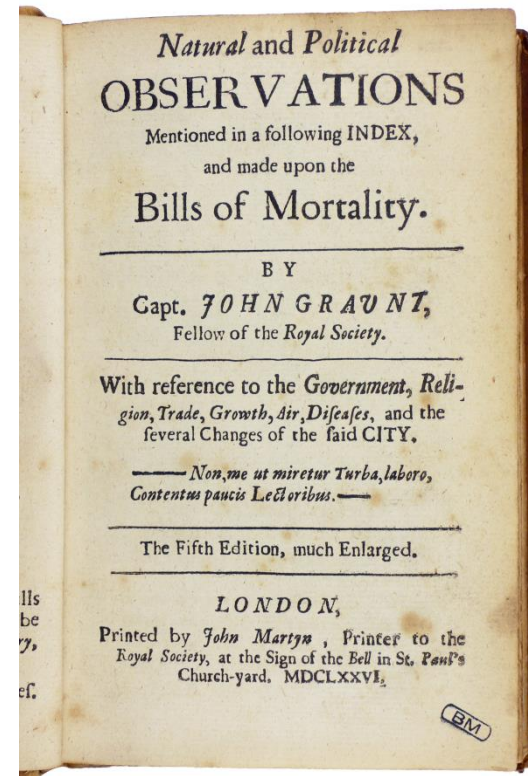
- Born London
- Haberdasher (Clothing and accessories)
- Early demographer and epidemiologist
 - Developed statistical and census methods
 - Derived early estimate of population of London and England



John Graunt

1620-1674

- *Natural and Political Observations Made Upon the Bills of Mortality* (1662)
 - Analysis of mortality in London by age, sex, cause
 - Primarily counts, fractions, but not rates
 - Five editions!
 - Charles II desired a system to warn of the onset and spread of bubonic plague.
 - System was never created
- Fellow Royal Society for the Improvement of Natural Knowledge
- Died of liver disease in London.



Blaise Pascal (1623–1662)

- French mathematician, physicist, religious philosopher
- Child prodigy educated by his father
- Invented a machine for addition and subtraction (age 15)
 - To help his father (a tax official)
- Contributed to
 - Study of fluids, pressure, vacuum
- Defended the scientific method
- Helped create
 - Solid geometry (age 16)
 - Probability theory with Pierre de Fermat
 - **Problem of points**, (problem of **division of the stakes**)
- Influenced development
 - Economics, social science
- Following a miraculous cure of his niece (1654) devoted to theology and philosophy



Pierre de Fermat

1601 (or 1607) - 1665

- French lawyer, mathematician
- Contributed to birth of
 - Differential calculus
 - Greatest and smallest ordinates of curved lines
 - Analytic geometry
 - Probability
 - Problem of points, (problem of division of the stakes)
 - Worked with Pascal
 - Optics



John Snow

1813-1858

- British physician
- Father of epidemiology
 - *On the Mode of Communication of Cholera* (1849)
 - Criticized **miasma theory**
 - 1855 edition: Investigation of epidemic in Soho, 1854
 - Broad Street pump (now Broadwick)
- Calculated dosage of ether and chloroform
 - Anesthetized Queen Victoria during labor



Miasmatic Theory of Disease

- Disease
 - Cholera, Clamydia, Black Death, etc.
- Caused by *miasma* from rotting organic matter.
 - (Μίασμα "pollution"), "bad air"
- Ancient Europe, India, and China.
- Theory abandoned after 1880

John Snow

Letter to the Editor of the Medical Times and Gazette

. . . , I found that nearly all the deaths had taken place within a short distance of the [Broad Street] pump. There were only ten deaths in houses situated decidedly nearer to another street-pump. In five of these cases the families of the deceased persons informed me that they always sent to the pump in Broad Street, as they preferred the water to that of the pumps which were nearer. In three other cases, the deceased were children who went to school near the pump in Broad Street...

John Snow

Letter to the Editor of the Medical Times and Gazette

. . . , I found that nearly all the deaths had taken place within a short distance of the [Broad Street] pump. There were only **ten deaths** in houses situated decidedly nearer to another street-pump. In **five** of these cases the families of the deceased persons informed me that they always sent to the pump in Broad Street, as they preferred the water to that of the pumps which were nearer. In **three** other cases, the deceased were children who went to school near the pump in Broad Street...

Explained 8 of 10 death not near pump!

John Snow

Letter to the Editor of the Medical Times and Gazette

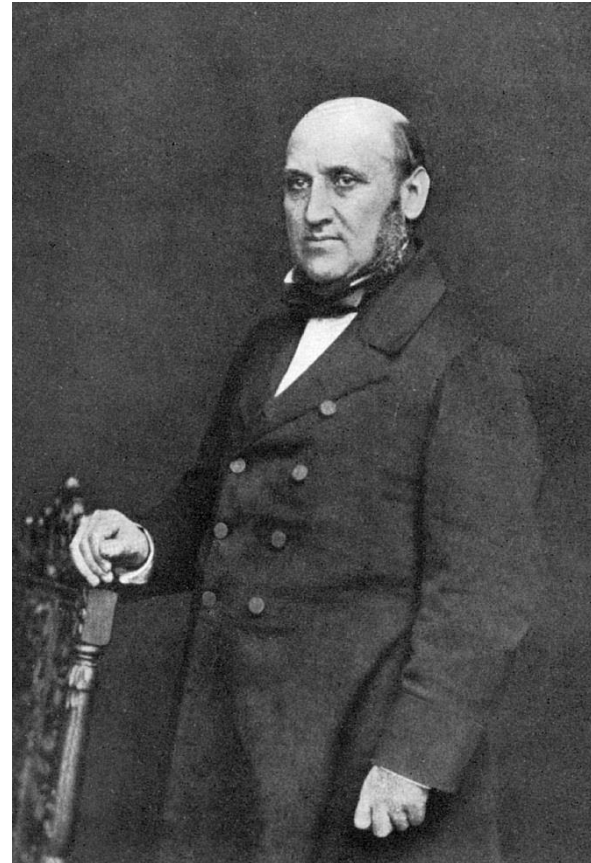
With regard to the deaths occurring in the locality belonging to the pump, there were 61 instances in which I was informed that the deceased persons used to drink the pump water from Broad Street, either constantly or occasionally...

The result of the inquiry, then, is, that there has been no particular outbreak or prevalence of cholera in this part of London except among the persons who were in the habit of drinking the water of the above-mentioned pump well.

I had an interview with the Board of Guardians of St James's parish, on the evening of the 7th inst [Sept 7], and represented the above circumstances to them. In consequence of what I said, the handle of the pump was removed on the following day.

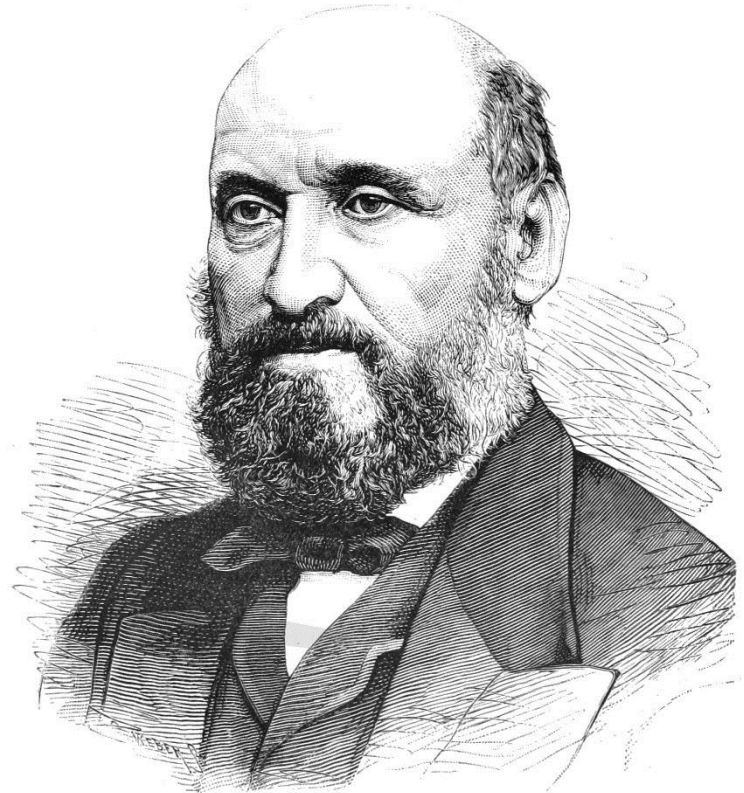
Dr. William Farr

- Assistant commissioner, London census 1851
- Supported miasma theory
 - Cholera transmitted by air
 - *Miasmata* along River Thames banks



Dr. William Farr

- Farr's miasma overshadowed Snow's theory
 - Cholera spread through water
 - Delayed response to outbreaks in Soho district of London and other areas



Cholera in London

- Outbreak in 1849
 - Killed ~15,000
- Second epidemic in 1853
 - Studied by Farr and Snow
 - Statistical evidence.

Cholera in London

- Snow:
 - Southwark & Vauxhall, and Lambeth water companies
 - Water directly from the Thames River
 - Implicated water source
- Farr:
 - Part of the General Board of Health 1854 Committee for Scientific Enquiries.
 - Farr showed an inverse correlation of mortality and elevation.
 - Consistent with miasma

Cholera in London

- Snow's theory not accepted, though evidence was taken seriously.

Bravo William Farr!

- Third Epidemic in 1866
 - Snow had died
 - **Farr accepted Snow's theory**
 - Monograph: high mortality people who drew their water from the Old Ford Reservoir in East London.
 - Farr's work considered conclusive.

Statistical Inference

How Do We Gain New Knowledge

- Make an observation
- Develop a theory explaining the observation
- Design an experiment to test the theory
- Come to a conclusion

Designing an Experiment

- Develop a testable hypothesis
 - H_0 :
- Develop an alternate hypothesis
 - H_a :
- H_0 and H_a
 - Mutually exclusive
 - Exhaustive

Carry Out Your Experiment

- Express the results of your experiment as a p value
 - Probability of finding by chance an outcome as extreme, or more extreme than the one you found
- $P \leq 0.05$ reject H_0 , accept H_a
- $P > 0.05$ accept H_0

Applies if you are a Frequentist, Bayesians use slightly different methods. . .

Construct 95% Confidence Interval

- 95% CI around a parameter
 $\beta - (1.96 \cdot SE_{\beta})$ to $\beta + (1.96 \cdot SE_{\beta})$

Construct 95% Confidence Interval

- 95% CI around a parameter

$$\beta - (1.96 \cdot SE_{\beta}) \text{ to } \beta + (1.96 \cdot SE_{\beta})$$

Assume

$$\beta = 40, SE_{\beta} = 2$$

$$40 - (1.96 \cdot 2) \text{ to } 40 + (1.96 \cdot 2)$$

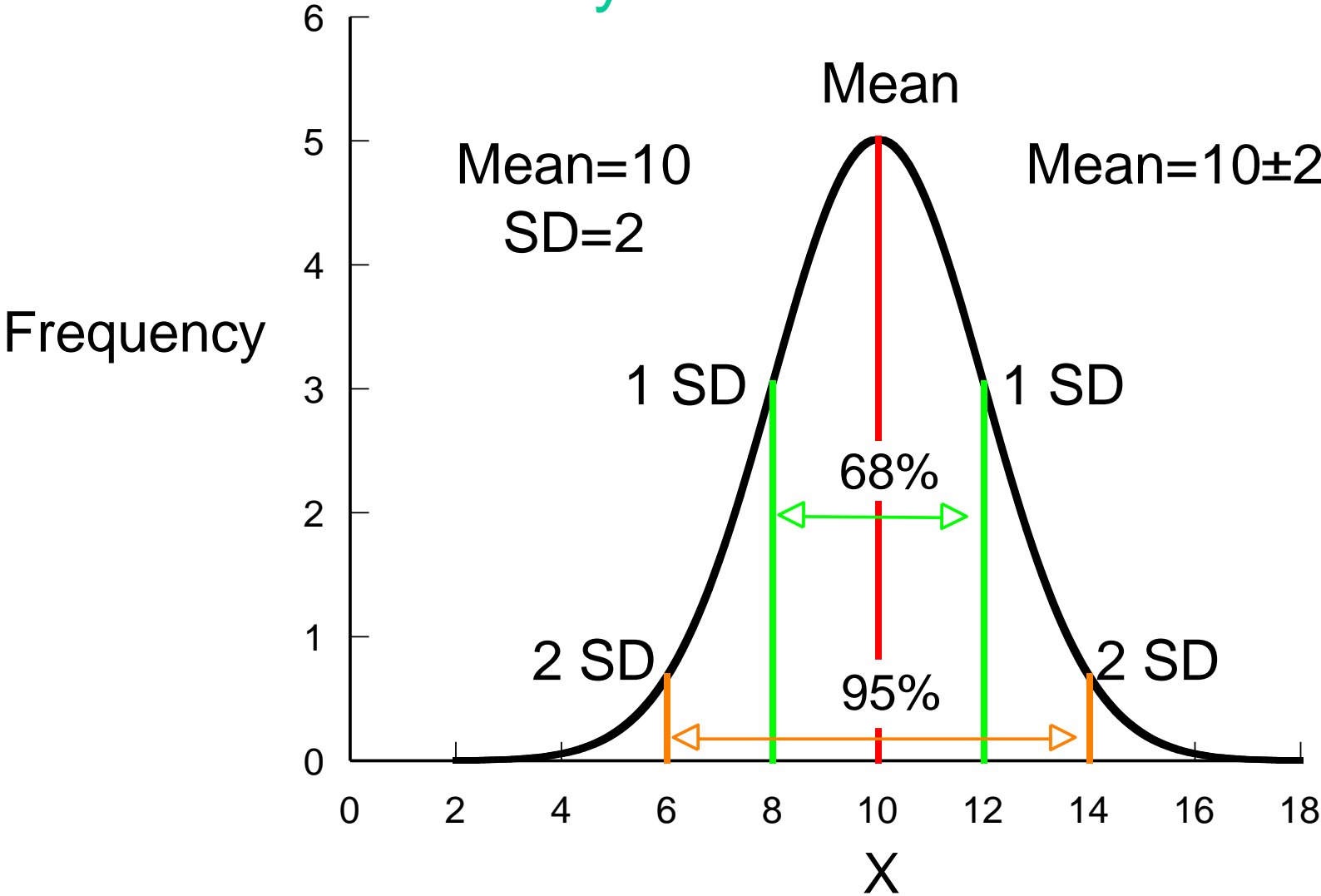
$$40 - (4) \text{ to } 40 + (4)$$

$$CI = 36 \text{ to } 44$$

Interpret the interval!

Construct 95% Confidence Interval

Why Use 1.96



Observation

- Most men I see appear to be taller than most women I see.

Hypothesis

- H_0 : There is no difference in height between men and women.
 - (Men and women are of the same height)
- H_a : Men and women are of different height.

Experiment

- Among all students at the University of Ibaden
 - Randomly select 10 men and 10 women
 - Measure height (without shoes) between 8 AM and 9 AM

Results

Men	Women
174.3	162.9
184.5	189.3
176.6	164.2
178.8	178.7
174.8	170.7
174.3	161.0
189.1	172.6
185.4	176.7
180.2	158.5
182.0	183.8

- $t = \frac{180.0 - 171.8}{\sqrt{1.63^2 + 3.26^2}}$
- $t = \frac{8.2}{3.64} = 2.24$
- $p_{t2.24, df18} < 0.04$

Mean	180.0	171.8
SD	5.17	10.30
SE	1.63	3.26

Conclusion

- Mean height difference
8.2 cm, $p < 0.04$; 95% CI (0.3-16.0)

Conclusion

- Mean height difference
8.2 cm, $p < 0.04$; 95% CI (0.3-16.0)
- Reject H_0
 - H_0 : There is no difference in height between men and women.
- Accept H_a
 - H_a : Men and women are of different height.
- Conclude there is evidence that men are taller than women
- **Statistics underlies the entire procedure!**

Statistical Jargon

Statistical Terms

- Independent Variable
- Dependent Variable
- Null Hypothesis
- Alternate Hypothesis
- P value
 - Significant
 - Non-significant
- 95% Confidence Interval
- Controlling for a variable

Statistical Terms

- **Independent Variable**
- Dependent Variable
- Null Hypothesis
- Alternate Hypothesis
- P value
 - Significant
 - Non-significant
- 95% Confidence Interval
- Controlling for a variable

Independent Variable

- Variables that are under the control of, or varied by the investigator
 - Study of an enzymatic reaction
 - Temperature
 - Substrate concentration, etc.

Independent Variable

- Variables that are thought to influence the outcome being studied
 - Study of heart disease incidence
 - Age
 - Sex
 - Cholesterol

Statistical Terms

- Independent Variable
- **Dependent Variable**
- Null Hypothesis
- Alternate Hypothesis
- P value
 - Significant
 - Non-significant
- 95% Confidence Interval
- Controlling for a variable

Dependent Variable

- The outcome being studied
- Variable that responds to experimental manipulation
 - Study of an enzymatic reaction
 - Yield
 - Study of heart disease incidence
 - Number of new cases

Statistical Terms

- Independent Variable
- Dependent Variable
- **Null Hypothesis**
- Alternate Hypothesis
- P value
 - Significant
 - Non-significant
- 95% Confidence Interval
- Controlling for a variable

Null Hypothesis

- A statement that we try to disprove
 - A “straw” man
 - Sometimes referred to as H_0

Null Hypothesis

- A statement that we try to disprove
 - A “straw” man
 - Sometimes referred to as H_0
- Stated so that disproving the null leads to an unambiguous conclusion
 - H_0 : Treatment does not change CD4 count

A Comparison of two Means

Null Hypothesis

Group 1		Group 2	
Mean	SE	Mean	SE
40	10	80	30

- H0: The groups come from the same population (and thus have the same mean save for sampling variation).

Statistical Terms

- Independent Variable
- Dependent Variable
- Null Hypothesis
- **Alternate Hypothesis**
- P value
 - Significant
 - Non-significant
- 95% Confidence Interval
- Controlling for a variable

Alternate Hypothesis

- A statement we accept if we disprove the null hypothesis.
 - Sometimes referred to as H_a
- H_0 and H_a
 - Mutually exclusive
 - If one is true, the other must be false (and conversely)
 - Collectively exhaustive
 - Must subsume all possible conditions.

Alternate and Null Hypothesis

Mutually Exclusive, Exhaustive

- Mutually **exclusive**
 - If one is true, the other must be false (and conversely)
 - H₀: John is **alive**
 - H_a: John is **dead**
- Collectively **exhaustive**
 - Must subsume **all possible conditions**.
 - H₀: John is **alive**
 - H_a: John is **dead**

Alternate and Null Hypothesis

Mutually Exclusive, Exhaustive

- Not Mutually exclusive
 - If one is true, the other may or may not be false (and conversely)
 - H₀: John is talking
 - H_a: John is walking
- Not Collectively Exhaustive
 - Do not subsume all possible conditions.
 - H₀: John is running
 - H_a: John is walking
 - H_a: John is standing still

A Comparison of two Means

Null Hypothesis, H0

Group 1		Group 2	
Mean	SE	Mean	SE
40	10	80	30

- H0: The groups come from the same population and thus should have the same mean (save for sampling variation).

A Comparison of two Means

Alternate Hypothesis, H_a

Group 1		Group 2	
Mean	SE	Mean	SE
40	10	80	30

- H_0 : The groups come from the same population and thus should have the same mean (save for sampling variation).
- H_a : The groups come from different populations and thus have different means

Statistical Terms

- Independent Variable
- Dependent Variable
- Null Hypothesis
- Alternate Hypothesis
- **P value**
 - Significant
 - Non-significant
- 95% Confidence Interval
- Controlling for a variable
-

P value

- The probability of getting, purely by chance, a result
 - as extreme or more extreme than the one observed

A Comparison of two Means

(20 subjects in each group)

Group 1		Group 2	
Mean	SE	Mean	SE
40	10	80	30

A Comparison of two Means

Null Hypothesis, H0

Group 1		Group 2	
Mean	SE	Mean	SE
40	10	80	30

- H0: The groups come from the same population (and thus have the same mean save for sampling variation).

A Comparison of two Means

Alternate Hypothesis, H_a

Group 1		Group 2	
Mean	SE	Mean	SE
40	10	80	30

- H_0 : The groups come from the same population (and thus have the same mean save for sampling variation).
- H_a : The groups come from different populations and thus have different means.

A Comparison of two Means

Student's t-Test

Group 1		Group 2	
Mean	SE	Mean	SE
40	10	80	30

- H0: The groups come from the same population (and thus have the same mean save for sampling variation).

$$t = \frac{\text{mean}_2 - \text{mean}_1}{SE_{\text{Difference}}}$$

A Comparison of two Means

Student's t-Test

Group 1		Group 2	
Mean	SE	Mean	SE
40	10	80	30

- H0: The groups come from the same population (and thus have the same mean save for sampling variation).

$$t = \frac{\text{mean}_2 - \text{mean}_1}{SE_{\text{Difference}}}$$

$$t = \frac{80 - 40}{\sqrt{10^2 + 30^2}} = \frac{40}{32} = 1.3$$

A Comparison of two Means

Student's t-Test

Group 1		Group 2	
Mean	SE	Mean	SE
40	10	80	30

- H0: The groups come from the same population (and thus have the same mean save for sampling variation).

$$t = \frac{mean_2 - mean_1}{SE_{Difference}} \qquad t = \frac{80 - 40}{32} = 1.3$$

$$p(t_{1.3,38}) = 0.21$$

Who Invented Student's t-Test?

Student's t-Test

William Gossett 1876-1937



- Chief Statistician (quality control inspector) for Guinness brewery in Dublin
- Employer's regulations concerning trade secrets prevented him from publishing his discovery. Because of the importance of the t distribution, Gossett was allowed to publish under the pseudonym "Student".

BIOMETRIKA.

THE PROBABLE ERROR OF A MEAN.

BY STUDENT.

Introduction.

ANY experiment may be regarded as forming an individual of a "population" of experiments which might be performed under the same conditions. A series of experiments is a sample drawn from this population.

Now any series of experiments is only of value in so far as it enables us to form a judgment as to the statistical constants of the population to which the experiments belong. In a great number of cases the question finally turns on the value of a mean, either directly, or as the mean difference between the two quantities.

A Comparison of two Means

P Values

Group 1		Group 2		Difference		t	p
Mean	SE	Mean	SE	Mean	SE		
40	10	80	30	40	32	1.26	0.21

- If H_0 is true
 - The chance of obtaining two means differing by $80-40=40$ (or more) by chance is 21 in 100 or 21%

A Comparison of two Means

P Values

Group 1		Group 2		Difference		t	p
Mean	SE	Mean	SE	Mean	SE		
40	10	80	30	40	32	1.26	0.21
40	10	90	30	50	32	1.58	0.12

- If H_0 is true
 - The chance of obtaining two means differing by $90-40=50$ (or more) by chance is 12 in 100 or 12%

A Comparison of two Means

P Values

Group 1		Group 2		Difference		t	p
Mean	SE	Mean	SE	Mean	SE		
40	10	80	30	40	32	1.26	0.21
40	10	90	30	50	32	1.58	0.12
40	10	100	30	60	32	1.90	0.07

- If H_0 is true
 - The chance of obtaining two means differing by $100-40=60$ (or more) by chance is 7 in 100 or 7%

A Comparison of two Means

P Values

Group 1		Group 2		Difference		t	p
Mean	SE	Mean	SE	Mean	SE		
40	10	80	30	40	32	1.26	0.21
40	10	90	30	50	32	1.58	0.12
40	10	100	30	60	32	1.90	0.07
40	10	105	30	65	32	2.06	0.05

- If H_0 is true
 - The chance of obtaining two means differing by $105-40=65$ (or more) by chance is 5 in 100 or 5%

Statistical Terms

- Independent Variable
- Dependent Variable
- Null Hypothesis
- P value
 - **Significant**
 - Non-significant
- 95% Confidence Interval

Statistically Significant

- A finding so unlikely to be found solely by chance that we accept it as being “true”
 - (Informing us about the world)
- P small enough to reject the null hypothesis
 - $P \leq 0.05$
 - 5 chances in 100 or
 - 1 chance in 20
 - Of finding, by chance,
 - a value as far or farther from the null hypothesis as the one observed.

A Significant Result

$$p \leq 0.05$$

Group 1		Group 2		Difference		t	p
Mean	SE	Mean	SE	Mean	SE		
40	10	80	30	40	32	1.26	0.21
40	10	90	30	50	32	1.58	0.12
40	10	100	30	60	32	1.90	0.07
40	10	105	30	65	32	2.06	0.05

- We **reject the null hypothesis** if the difference between the means **equal to or greater than 65**.
 - We say the samples do not come from the same population (and thus can have different means)

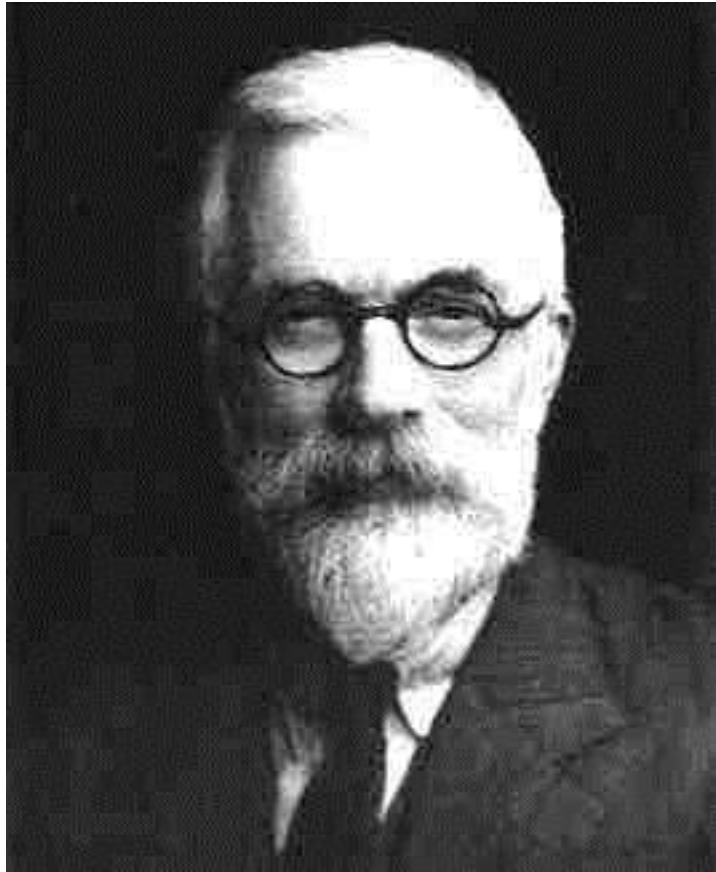
Whence Comes $p \leq 0.05$?

Whence Comes $p \leq 0.05$?

Is it evidence-based?

Where does $p \leq 0.05$ come from?

Sir Ronald Aylmer Fisher 1890-1962



- Statistician at Rothamsted Agricultural Experimental Station
- Introduced concepts of $p < 0.05$, randomization, analysis of variance, and likelihood

Where does $p \leq 0.05$ come from?

R.A. Fisher

"If one in twenty does not seem high enough odds, we may, if we prefer it, draw the line at one in fifty (the 2 percent point), or one in a hundred (the 1 percent point). Personally, the writer prefers to set a low standard of significance at the 5 percent point, and ignore entirely all results which fail to reach that level. A scientific fact should be regarded as experimentally established only if a properly designed experiment rarely fails to give this level of significance."

Fisher, R. A. (1926), "The arrangement of field experiments," Journal of the Ministry of Agriculture of Great Britain, Volume 33, p. 504.

Statistical Terms

- Independent Variable
- Dependent Variable
- Null Hypothesis
- P value
 - Significant
 - **Non-significant**
- 95% Confidence Interval

Non-Significant

- A result sufficiently likely to have occurred solely due to chance that we do not accept it as being “true” (not informing us about the world)
- P too large to reject the null hypothesis
 - $P > 0.05$
 - Greater than 5 chances in 100 or
 - 1 chance in 20
 - Of finding, by chance,
 - a value as far or farther from the null hypothesis as the one observed.

A Non-Significant Result

$p > 0.05$

Group 1		Group 2		Difference		t	p
Mean	SE	Mean	SE	Mean	SE		
40	10	80	30	40	32	1.26	0.21
40	10	90	30	50	32	1.58	0.12
40	10	100	30	60	32	1.90	0.07
40	10	105	30	65	32	2.06	0.05

- We **accept the null hypothesis** if the difference between the means **less than 65**.
 - We have no evidence that samples came from different populations
 - We have no evidence that the means of the samples are different.

A Significant Result

$$p < 0.05$$

- Reasons **for** rejecting H0
 - Saying we have a significant effect
 - The means of the two groups are different
 - Random variation (chance)
 - Bias
 - Confounding

A Significant Result

$$p < 0.05$$

- Reasons **for** rejecting H_0
 - Saying we have a significant effect
 - The means of the two groups are different
 - Random variation (chance)
 - Bias
 - Any error in the design, conduct, analysis, or reporting of a study that leads to an incorrect conclusion.
 - Confounding

A Significant Result

$$p \leq 0.05$$

- Reasons **for rejecting** H0
 - Saying we have a significant effect
 - The means of the two groups are different
 - Random variation (chance)
 - Bias
 - Confounding
 - An error caused by failing to take into consideration factors that influence both outcome and exposure

A Non-Significant Result

$$p > 0.05$$

- Reasons for **not rejecting** H0
 - Saying our results are not significant
 - The means of the two groups are the same
 - We do not have enough “power” (i.e. our sample size is too small) to demonstrate a difference
 - Confounding
 - Random error (chance)
 - Bias

A Non-Significant Result

What Does a Non-Significant Mean?

- Failure to reject the null hypothesis
 - (a non-significant result)
 - Does not mean the means are be same, but rather
 - We have no evidence that they are different!

P Value Truisms

- A p of 0.05 (a significant value)
 - is not different from
 - a p of 0.06 (a non-significant value).
- Statistical significance is not the same as biological importance.
 - A small p value (e.g. 0.0001) does not indicate an important result.

P Value Truisms

- No p value, no matter how small, establishes **absolute** truth
- RE-search
 - Old French: re-re+ cercher to search

P Values

- When reading a paper
 - Are the results
 - Statistically significant?
 - Biologically important?
 - If not significant
 - Do they suggest an important finding?
 - Have other researchers found similar results?

Statistical Terms

- Independent Variable
- Dependent Variable
- Null Hypothesis
- Alternate Hypothesis
- P value
 - Significant
 - Non-significant
- **95% Confidence Interval**
- Controlling for a variable

95% Confidence Interval (95% CI)

- A range of numbers constructed so that there is a 95% probability that the value of the parameter measured will fall within the bounds of the interval.

95% Confidence Interval (95% CI)

- A range of numbers constructed so that there is a 95% probability that the value of the parameter measured will fall within the bounds of the interval.
- A range of values which will include the results of repeated experiments 95 out of 100 times.

95% Confidence Interval

Group 1		Group 2		Difference	
Mean	SE	Mean	SE	Value	95% CI
40	10	80	30	40	-22 to 102
40	10	90	30	50	-12 to 112
40	10	100	30	60	-2 to 122
40	10	105	30	65	3 to 127

- Depending on the analysis
 - A 95% CI that **excludes zero** implies significance
 - Differences
 - A 95% CI that **excludes one** implies significance
 - Ratios

Confidence Intervals

- 95% most commonly used
- Other intervals can be computed
 - 99%
 - 97.5%
 - 90%

Confidence Intervals

- When reading a paper
 - Are the intervals wide or small?
 - Do the intervals include 0?
 - Student's t test
 - Regression
 - Correlation
 - Chi square
 - Do the intervals include 1?
 - Relative risk
 - Odds ratio
- Use **zero** for statistics that are based on a **difference**
- Use **one** for statistics that are based on a **quotient**

P value vs. Confidence Interval

- P values
 - Compact
 - Easy to communicate
 - Universally “understood”
 - Come with “dogmatic” baggage
 - $P < 0.05$ is significant
 - Completely arbitrary
- Confidence Intervals
 - Less compact
 - Convey more information
 - Not Universally understood
 - Come with less baggage

Statistical Terms

- Independent Variable
- Dependent Variable
- Null Hypothesis
- Alternate Hypothesis
- P value
 - Significant
 - Non-significant
- 95% Confidence Interval
- **Controlling for a variable**

•

Controlling for a Variable

- When looking at the effect of one variable on another
 - Effect of x on y
- Controlling allows us to eliminate the potential confounding effect of a third variable
 - Effect on x on y controlling for z
- Allows us to get the true relation between x and y .

Controlling for a Variable

- Question: What is the relation between weight and blood pressure
- Need to control for height
 - Taller people weigh more than shorter people
 - Extra weight may not indicate greater obesity and may not effect blood pressure

$$\text{BloodPressure} = \beta_0 + \text{Weight}\beta_1 + \text{Height}\beta_2$$



Reubin and
Kris Andres'
Koi (Nishikigoi)
Pond

Scales on Which Data are Measured

Scales on Which Data Can be Measured

- Nominal
- Ordinal
- Interval
- Ratio

Nominal

- A scale where data are divided into categories
- There is no order to the categories
 - Sex
 - Male, Female
 - Vital Status
 - Dead, Alive
 - Ethnicity
 - Black, White, Hispanic, American Indian, Other
- Sometimes called a discrete or symbolic scale

Ordinal

- A scale of data organized into categories where there is a logical ordering to the categories.
 - Soft drink size
 - Small, medium, large
 - Pain scale
 - 0, 1, 2, 3, 4
- Differences (or other mathematical transformations) are ***not*** meaningful

Interval

- A scale where the difference between two values is interpretable
- Each step in the scale is of a fixed size
- Temperature in degrees Fahrenheit
- Equal difference implies equal distance
 - 100°F is 10 degrees greater than 90°F
- No "natural" zero
 - Ratios are meaningless.
 - 100°F is not twice as hot as 50 °F

Ratio

- A scale in which both differences and ratios are interpretable
 - 100 kg is twice as heavy as 50 kg
 - $100/50 = 2$
- Scale has a true zero
 - Temperature in degrees Kelvin
 - (0°K implies no molecular motion)
 - Weight (pounds, kilograms, etc.)

Ratio vs. Interval

- Distinction between interval and ratio data are subtle
 - Often not important.
- Certain specialized statistics can only be applied to ratio data.
 - Geometric mean
 - Coefficient of variation
- True interval data are rare

Information Content of Scales

- Nominal
- Ordinal
- Interval
- Ratio



Information
Content
Increases

Why is the Scale Important?

Scale can Dictate Statistic

- Nominal data
 - Mean, median, and standard deviation have no meaning
 - Mean or SD for race or ethnicity?
- Ordinal data
 - Computation of a median can be justified,
 - Some statisticians question computing the mean

Ordinal Data

Median and Mean

- Median interpretable
- Mean?
 - Pain scale
 - 0, 1, 2, 3, 4
 - 0, 1, . . . 2, 3, 4
 - Median 2

$$\text{Mean} = \frac{0+1+2+3+4}{5} = 2$$

Interval and Ratio Data

- Mean, median, standard deviation have meaning

Why is the Scale Important?

Scale can Dictate Statistic

- Different statistical methods are used for data measured on different scales
 - Nominal and Ordinal
 - Vs.
 - Interval and Ratio

Statistics: Nominal Data

- Logistic regression
- Log-linear regression
- Independent Samples
 - Pearson's Chi squared
 - Binomial test
 - Fisher's Exact test
- Related Samples
 - Fisher's Exact test
 - McNemar's chi-square test

Ordinal Data

Frequently Improperly Analyzed

- Warning!
 - We often use scales
 - CESD, Mini Mental, coronary calcium score, etc.
 - These scales are ORDINAL
 - Use **non-parametric** statistic!
 - Don't use
 - » Student's t-test
 - » Regression
 - Use
 - » **Mann-Whitney "U"**
 - » **Wilcoxon Rank Sum**

Parametric vs. Non-Parametric Statistics

- Parametric
- Non-Parametric

Parametric vs. Non-Parametric Statistics

- Parametric methods
 - Rely on assumptions about the underlying distribution of the data being analyzed
- Non-parametric methods
 - Make few (and less strong assumptions) about the underlying distribution of the data being analyzed

Parametric vs. Non-Parametric Statistics

- Parametric
 - Strong assumptions about data
 - E.g. normal distribution
 - If assumptions correct
 - More powerful
 - If distribution assumptions incorrect
 - Can give incorrect results
 - Able to adjust for multiple factors
- Non-parametric
 - Fewer, weaker assumptions about data
 - If assumptions correct
 - Slightly less powerful
 - If distribution assumptions incorrect
 - Can give correct results
 - Limited ability to adjust for multiple factors

Statistics: Ordinal Data

- Ordinal regression
- Spearman's rho (rank correlation coefficient)
- Independent Samples
 - Mann-Whitney U test
 - Median test
 - Kruskal-Wallis one-way ANOVA by ranks
 - Kolmogorov-Smirnov Test
 - Jonckheere's Trend
- Matched Samples
 - Sign Test
 - Wilcoxon test
 - Page's L

Parametric vs. Non-Parametric Statistics

- Parametric
 - An example
 - Student's t-test
- Non-Parametric

Student's t-Test

A Statistic Used for Interval or Ratio Data

- Outcome (dependent)
 - Interval or Ratio
- Predictor (independent)
 - Nominal – represents group membership
 - One or two groups
- Groups
 - Independent
 - Or
 - Related (two measurements from same subjects)
 - Paired t test

Student's t-Test

A Statistic Used for Interval or Ratio Data

- Limitations
 - Parametric method
 - Assumes outcome variable is normally distributed
 - For two groups:
 - Assumes variances in groups are equal
 - Small sample sizes (<10/group) or very different sample sizes can cause problems

Gedankenexperiment (Thought Experiment)

- Drug given to men and women
- Half-life measured

- Question:
- Does half life differ by sex?

Gedankenexperiment (Thought Experiment)

- What scale is used for
 - The outcome?
 - Half-life
 - The classification variable?
 - Sex

Gedankenexperiment (Thought Experiment)

- What scale is used for
 - The outcome?
 - Half-life - **Ratio**
 - The classification variable?
 - Sex - **Nominal**

Gedankenexperiment (Thought Experiment)

- What scale is used for
 - The outcome?
 - Half-life - **Ratio**
 - The classification variable?
 - Sex - **Nominal**
 - Statistic
 - **Student's t-test**

Gedankenexperiment Result

<u>Men</u>	<u>Women</u>
9	19
10	20
7	17
7	17
12	22
8	18
13	23
11	21
10	20
8	18

Gedankenexperiment Result

<u>Men</u>	<u>Women</u>
9	19
10	20
7	17
7	17
12	22
8	18
13	23
11	21
10	20
8	18

$$t = \frac{\bar{x} - \bar{y}}{SE(\bar{x} - \bar{y})}$$

			<u>t</u>	<u>df</u>	<u>p</u>
Mean	9.5	19.5	-10.81	18	<0.0001

Why is the Scale Important?

Scale can Dictate Statistic

- When reading a paper
 - What scales of measurement were used?
 - Do the statistical methods match the scale of measurement?

Why is the Scale Important?

Scale can Dictate Statistic

- Interval or Ratio data
 - Calculate means
 - Student's t test (paired or unpaired)
- Nominal or Ordinal data
 - Can't calculate mean
 - Other tests

Why is the Scale Important?

Scale can Dictate Statistic

Nominal data

- Data can be classified or grouped
 - Pearson's chi square
 - Odds ratio
 - Logistic regression

Why is the Scale Important?

Scale can Dictate Statistic

Ordinal data (unpaired data)

- Data can be ordered
 - **Mann-Whitney “U”**
 - Wilcoxon Rank Sum
 - Mann-Whitney-Wilcoxon
 - Median Test
 - **Kruskal-Wallis** one-way ANOVA by Ranks
 - Kolmogorov-Smirnov Test

Why is the Scale Important?

Scale can Dictate Statistic

Ordinal data (paired data)

- Data can be ordered
 - Sign Test
 - Wilcoxon Sign Rank Test



Uses of Statistics

Uses of Statistics

- To summarize data
 - Descriptive statistics
- To make inferences about data
 - Inferential statistics
 - Statistical inference

Uses of Statistics

- To summarize data
 - Descriptive statistics
- To make inferences about data
 - Inferential statistics
 - Statistical inference

Types of Statistics

- Inferential statistics
 - To make inferences about data
 - Is one batch of numbers larger than another?
 - Is there a relation between height and weight?
 - Are men stronger than women?

Uses of Statistics

Uses of Statistics

- To summarize data
 - Descriptive statistics
 - Where do “most“ of the value lie?
 - How variable are the values?
- To make inferences about data
 - Inferential statistics
 - Statistical inference

Descriptive Statistics

Summarization of Data

Location and Spread

- Measures of
 - **Location** (or central tendency)
 - Where do “most” of the data lie
 - **Spread** (or variability)
 - How much variability is there in the data

Summarization of Data

Location and Spread

- Descriptive statistics
 - Summarize data
 - Location
 - Mean
 - Median
 - Mode
 - Spread

Summarization of Data

Location and Spread

- Descriptive statistics
 - Summarize data
 - Location
 - Mean
 - Median
 - Mode
 - Spread
 - Variance
 - Standard deviation
 - Standard error

Measures of Location

- Mean: $\frac{\sum_{i=1}^n x_i}{n}$
- Median: “Middle” value
- Mode: Most “popular” value

Location

- Data: 0, 1, 2, 3, 4, 5, 5, 5, 6, 7, 8, 9, 10
- Mean: $(0+1+2+3+4+5+5+5+6+7+8+9+10)/13$
 $65/13 = 5$
- Median: 0,1,2,3,4,5,5,5,6,7,8,9,10
↑
5
- Mode: 0,1,2,3,4,5,5,5,6,7,8,9,10

Why have a Mean and Median?

- Mean is **sensitive**,
- Median is **resistant** to extreme values

- Data: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10
- Mean 5.5 Median $(5+6)/2=5.5$

- Data: 1, 2, 3, 4, 5, 6, 7, 8, 9, 1000
- Mean 104 Median $(5+6)/2=5.5$

Resistant vs. Sensitive Statistic

Resistant Statistic

- A statistic that is
 - relatively unchanged
 - when a large change is made
- in a small fraction of the data that are used to compute the statistic

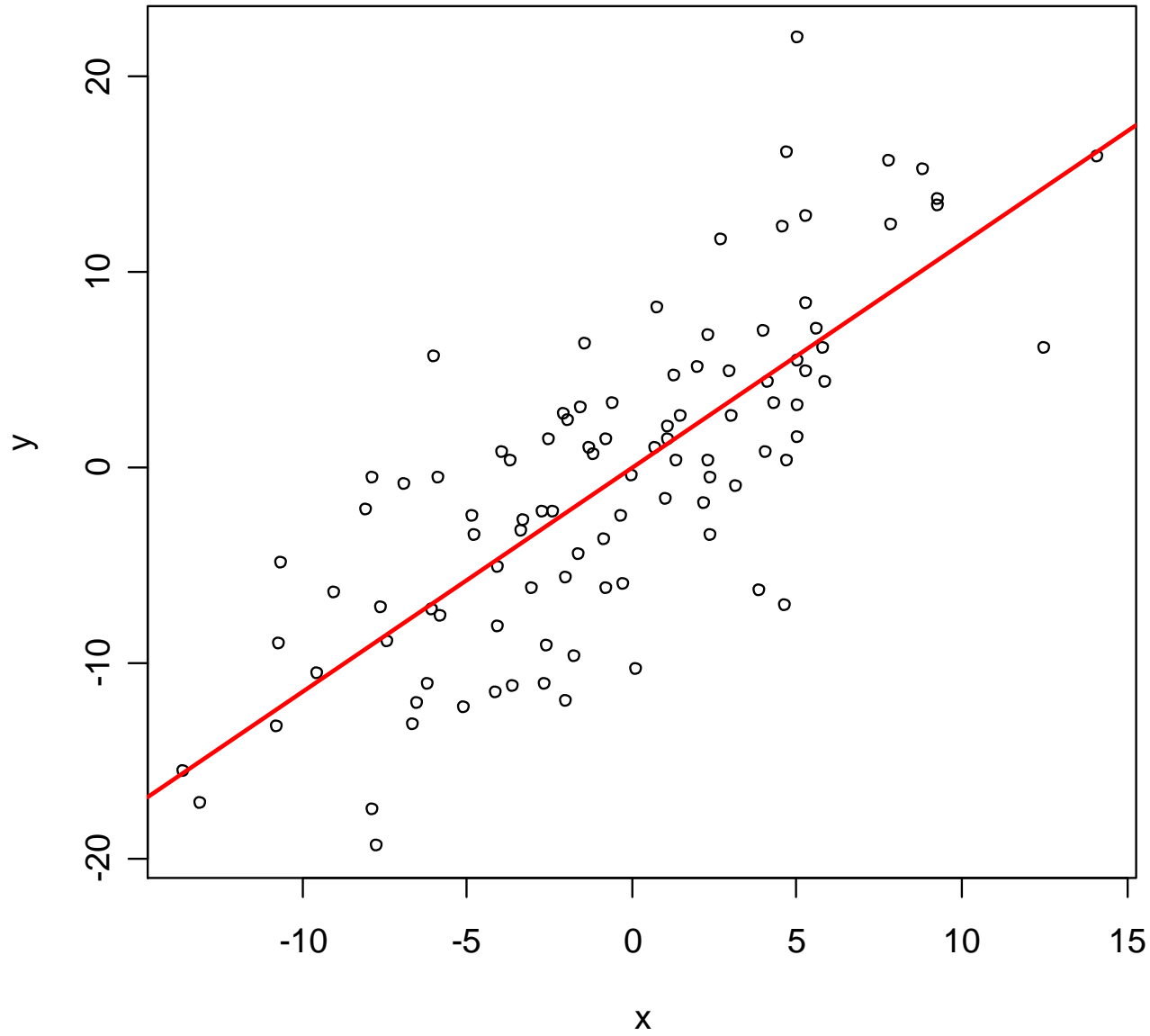
Sensitive Statistic

- A statistic that is
 - changed
 - when a small change is made
- in a small fraction of the data that are used to compute the statistic

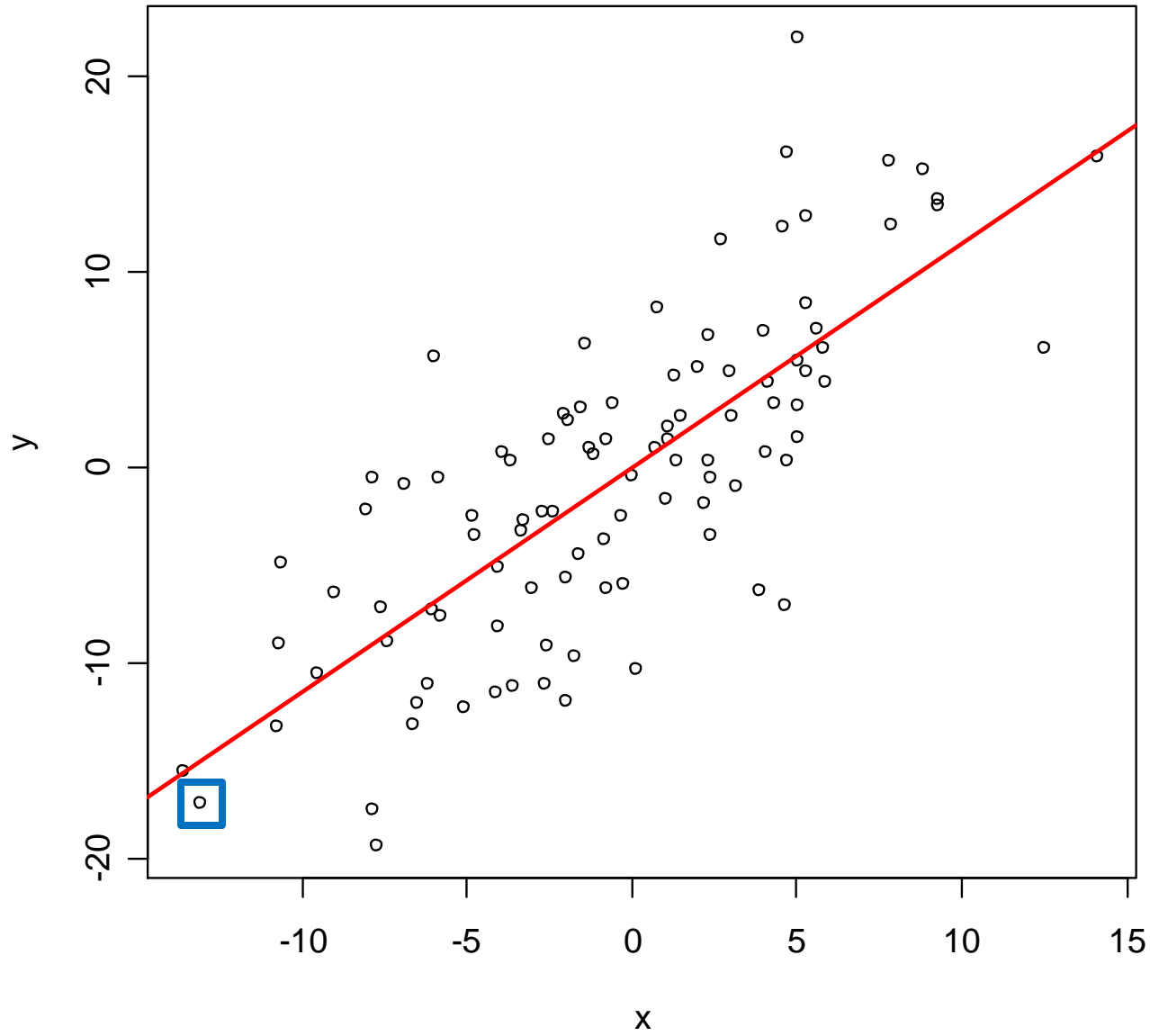
Robust Statistical Method

- A statistical method that produces results that are
 - relatively unchanged
 - when the data violate assumptions
 - (all statistical models have assumptions, e.g. data are normally distributed)

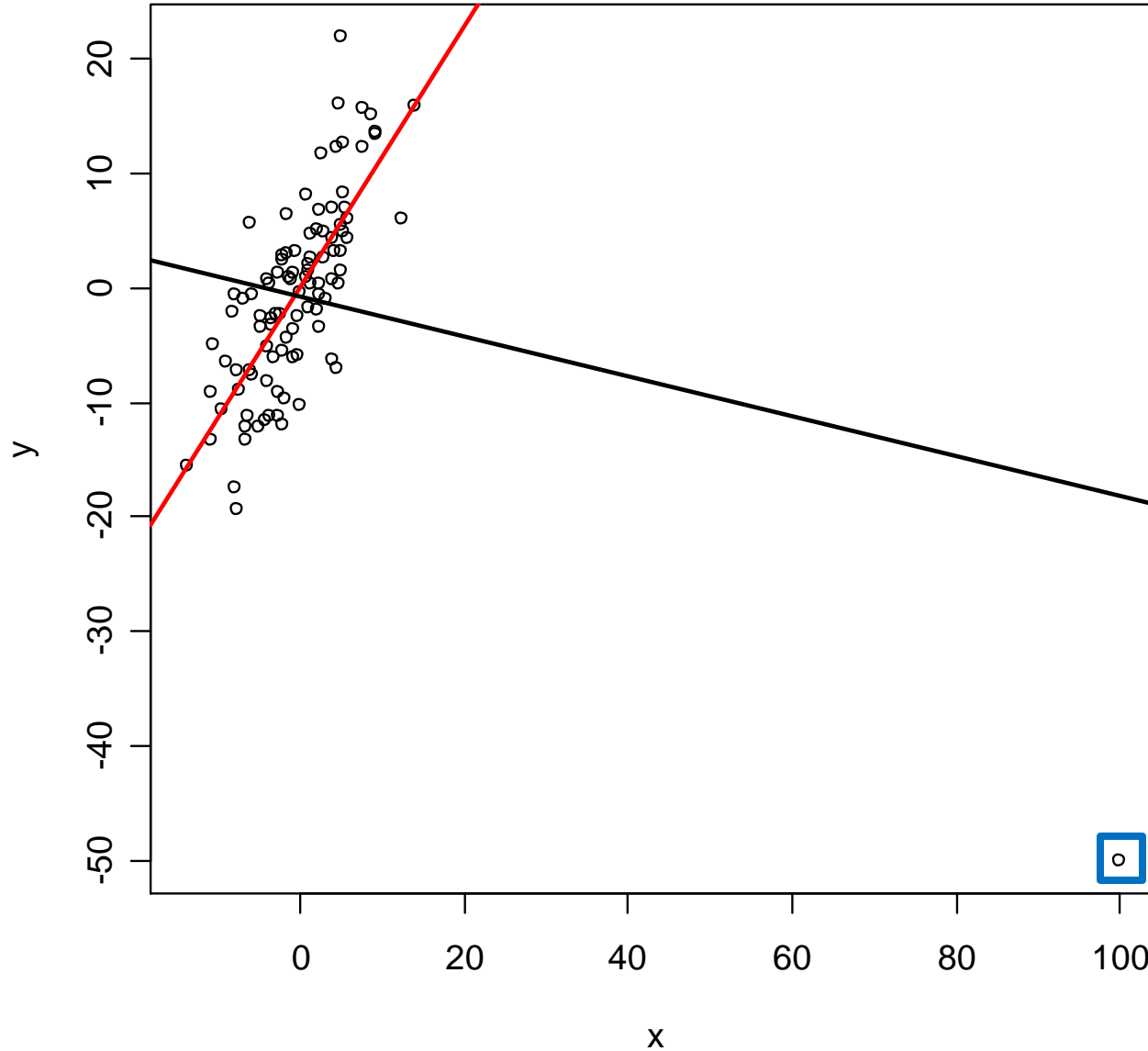
Regression is Not Robust



Regression is Not Robust

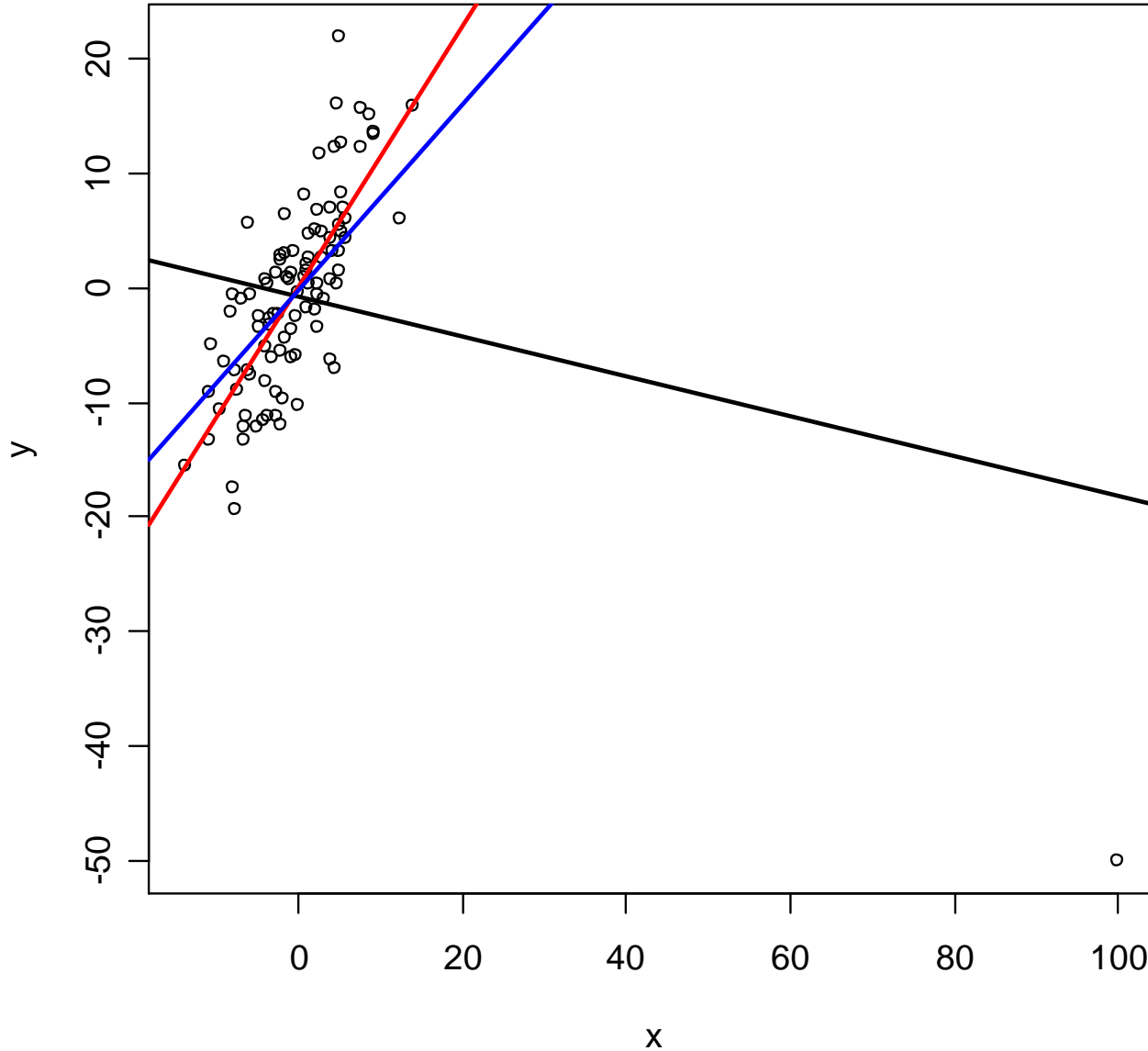


One Value Can Modify Fit



Single value
changed

Robust (Resistant) Regression



Lesson

- Many statistics are not resistant
- Many statistical methods are not robust
- Look at your data!
 - Know the assumptions of your method
 - Make sure the assumptions are met

Measures of Spread

- Variance
- Standard deviation
- Standard error
 - Quantify the variability of a measure

Measures of Spread - Variance

- Variance: The average of the squared distance from the mean

$$\sigma^2 = \sum_{i=1}^n (x_i - \bar{x})^2 / n \text{ or, } s^2 = \sum_{i=1}^n (x_i - \bar{x})^2 / (n-1)$$

- A measure of the spread of data whose scale is the square of the original data

Measures of Spread - Variance

- Variance: The average of the squared distance from the mean

$$\sigma^2 = \sum_{i=1}^n (x_i - \bar{x})^2 / n \text{ or, } s^2 = \sum_{i=1}^n (x_i - \bar{x})^2 / (n-1)$$

- Data: 1, 2, 3
- Mean= $(1+2+3)/3 = 2$
- Variance= $(1-2)^2 + (2-2)^2 + (3-2)^2 / (3-1) = 2/2 = 1$

Measures of Spread - Variance

- Variance: The average of the squared distance from the mean

$$\sigma^2 = \sum_{i=1}^n (x_i - \bar{x})^2 / n \text{ or, } s^2 = \sum_{i=1}^n (x_i - \bar{x})^2 / (n-1)$$

- Data: 1, 2, 3
- Mean= $(1+2+3)/3 = 2$
- Variance= $(1-2)^2 + (2-2)^2 + (3-2)^2 / (3-1) = 2/2 = 1$
- Units: The square of x
- e.g. x age in years, variance is in years²

Measures of Spread – Standard Deviation

- Standard deviation: Converts the variance back to its original scale of measurement

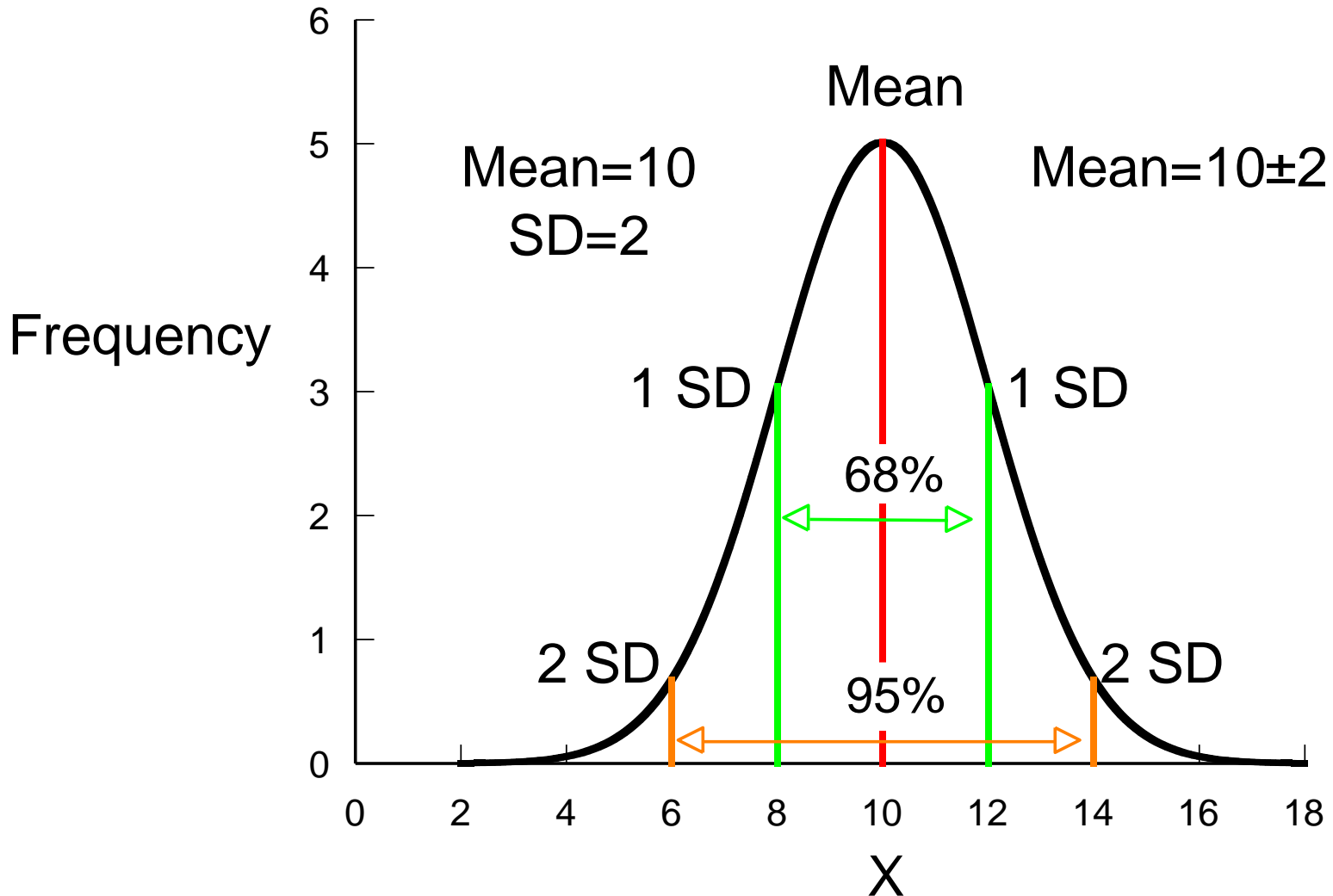
$$\sigma = \sqrt{\sigma^2} \quad SD = \sqrt{s^2}$$

- Units: The same as x
- e.g. x age in years,
- Variance is in years²,
- Standard deviation is in years

Standard Deviation

- When is the standard deviation used?
 - To *describe the distribution* of the data
- A measure of the **spread of data** measured on the same scale as the original data

Standard Deviation



Standard Deviation

100 Random Numbers, Mean \pm SD=99.7 \pm 21.9

40	44	64	65	65	66	66	67	67	68
69	69	70	73	75	78	79	81	82	82
83	84	84	84	84	84	85	85	86	86
86	87	87	90	91	92	92	93	93	94
95	95	96	97	98	98	99	99	100	102
102	103	103	103	105	106	106	106	107	107
107	107	108	108	109	109	110	110	111	112
112	112	113	113	113	114	114	115	115	117
119	120	121	121	122	123	124	124	126	128
131	131	132	133	133	134	135	137	150	150

Standard Deviation

100 Random Numbers, Mean \pm SD=99.7 \pm 21.9

40	44	64	65	65	66	66	67	67	68
69	69	70	73	75	78	79	81	82	82
83	84	84	84	84	84	85	85	86	86
86	87	87	90	91	92	92	93	93	94
95	95	96	97	98	98	99	99	100	102
102	103	103	103	105	106	106	106	107	107
107	107	108	108	109	109	110	110	111	112
112	112	113	113	113	114	114	115	115	117
119	120	121	121	122	123	124	124	126	128
131	131	132	133	133	134	135	137	150	150

Mean \pm 1 SD=78 to 122 (68% data)

Standard Deviation

100 Random Numbers, Mean \pm SD=99.7 \pm 21.9

40	44	64	65	65	66	66	67	67	68
69	69	70	73	75	78	79	81	82	82
83	84	84	84	84	84	85	85	86	86
86	87	87	90	91	92	92	93	93	94
95	95	96	97	98	98	99	99	100	102
102	103	103	103	105	106	106	106	107	107
107	107	108	108	109	109	110	110	111	112
112	112	113	113	113	114	114	115	115	117
119	120	121	121	122	123	124	124	126	128
131	131	132	133	133	134	135	137	150	150

Mean \pm 2 SD= **56 to 144 (95% data)**

Standard Deviation

100 Random Numbers

40	44	64	65	65	66	66	67	67	68
69	69	70	73	75	78	79	81	82	82
83	84	84	84	84	84	85	85	86	86
86	87	87	90	91	92	92	93	93	94
95	95	96	97	98	98	99	99	100	102
102	103	103	103	105	106	106	106	107	107
107	107	108	108	109	109	110	110	111	112
112	112	113	113	113	114	114	115	115	117
119	120	121	121	122	123	124	124	126	128
131	131	132	133	133	134	135	137	150	150

Mean \pm 3 SD= 34 to 165 (99% of data)

Standard Error

- Gives the distribution of sample means
- A measure of the **spread of mean** measured on the same scale as the original data
- Question: What is the concentration of sodium in the ocean?

Standard Error Calculation

- Experiment: Go to the Inner Harbor
- Do 10 times:
 - Draw a sea water sample
 - Record sodium concentration
 - Replace the sea water
- End
- Obtain the average of the 10 concentrations

Standard Error Calculation

- Is my sample representative of the entire ocean? . . . Go to Tahiti
- Do 10 times:
 - Draw a sea water sample
 - Record sodium concentration
 - Replace the sea water
- End
- Obtain the average of the 10 concentrations

Standard Error Calculation

- Is my sample representative of the entire ocean?. . . Go to Hawaii. . .
- Do 10 times:
 - Draw a sea water sample
 - Record sodium concentration
 - Replace the sea water
- End
- Obtain the average of the 10 concentrations

Measures of Spread – Standard Error

- Standard error: Same scale as standard deviation

$$SE = \sigma / \sqrt{n}$$

$$SE = \text{Standard deviation} / \sqrt{n}$$

- Units: The same as x
- e.g. x age in years,
- Variance is in years²,
- Standard error is in years

Standard Error

Relation to Standard Deviation

- A **histogram** of the **means** will follow a **normal distribution**
 - (Central Limit Theorem)
- The standard deviation of the means is the standard error of the mean
 - $SE = SD(\text{mean}_1, \text{mean}_2, \dots, \text{mean}_n)$

Standard Error

Relation to Standard Deviation

- A histogram of the means will follow a normal distribution
 - (Central Limit Theorem)
- The **standard deviation** of the **means** is the **standard error of the mean**
 - $SE = SD(\text{mean}_1, \text{mean}_2, \dots, \text{mean}_n)$

Standard Error

Relation to Standard Deviation

Salt concentration of salt* in sea water (g/kg)

Measurement	Location		
	Inner Harbor	Tahiti	Hawaii
1	33.4	37.7	36.3
2	33.4	34.3	35.6
3	32.2	34.1	33.3
4	33.0	35.1	36.1
5	33.4	36.6	33.2
6	32.5	35.8	34.3
7	32.1	34.9	36.9
8	33.2	36.4	37.2
9	32.3	34.7	33.4
10	33.1	35.6	33.4
Mean	32.9	35.5	35.0

* Mostly NaCl

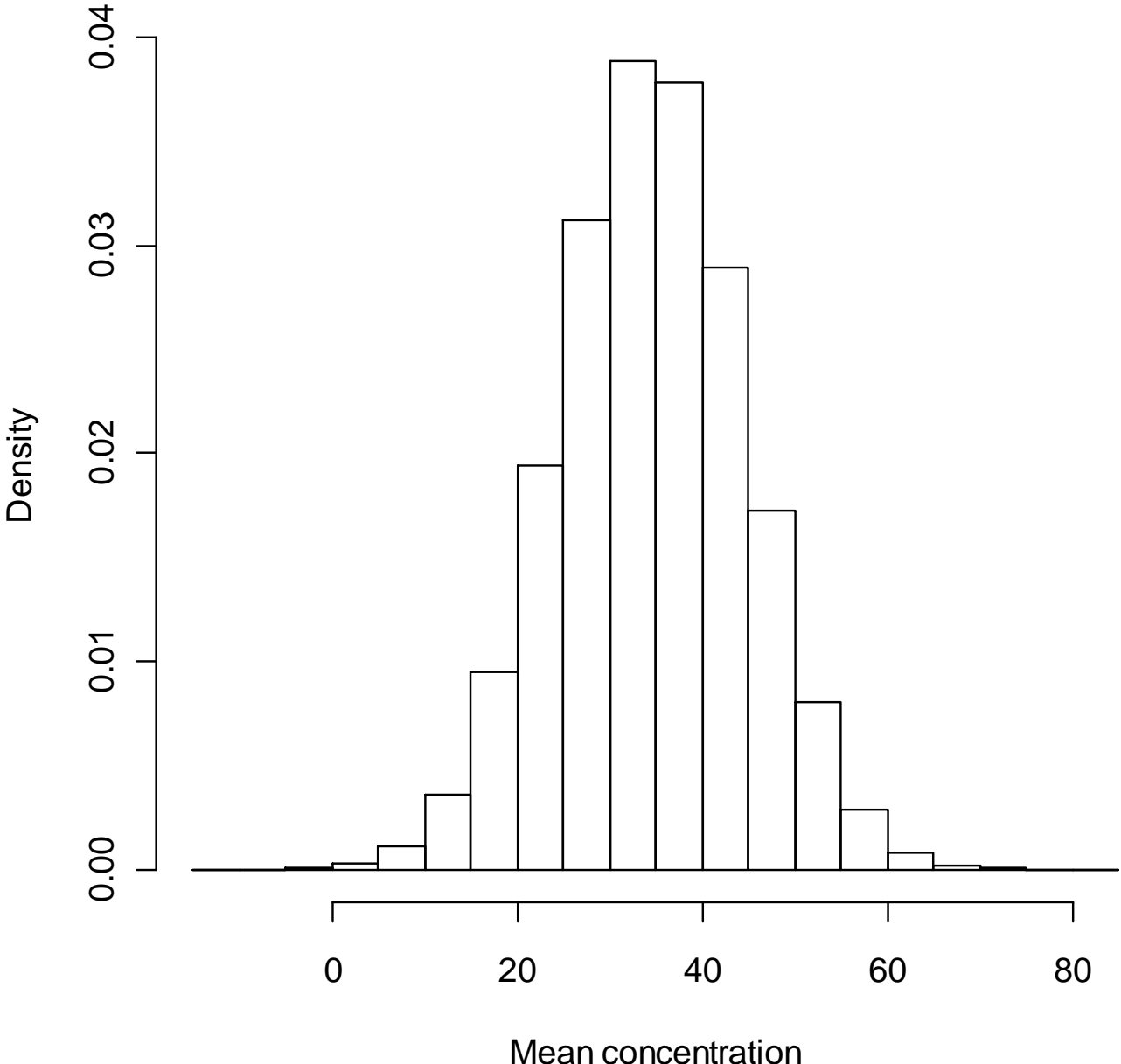
- $SE = SD(32.9, 35.5, \dots, 35.0)$

What is the salinity of Ocean Water?

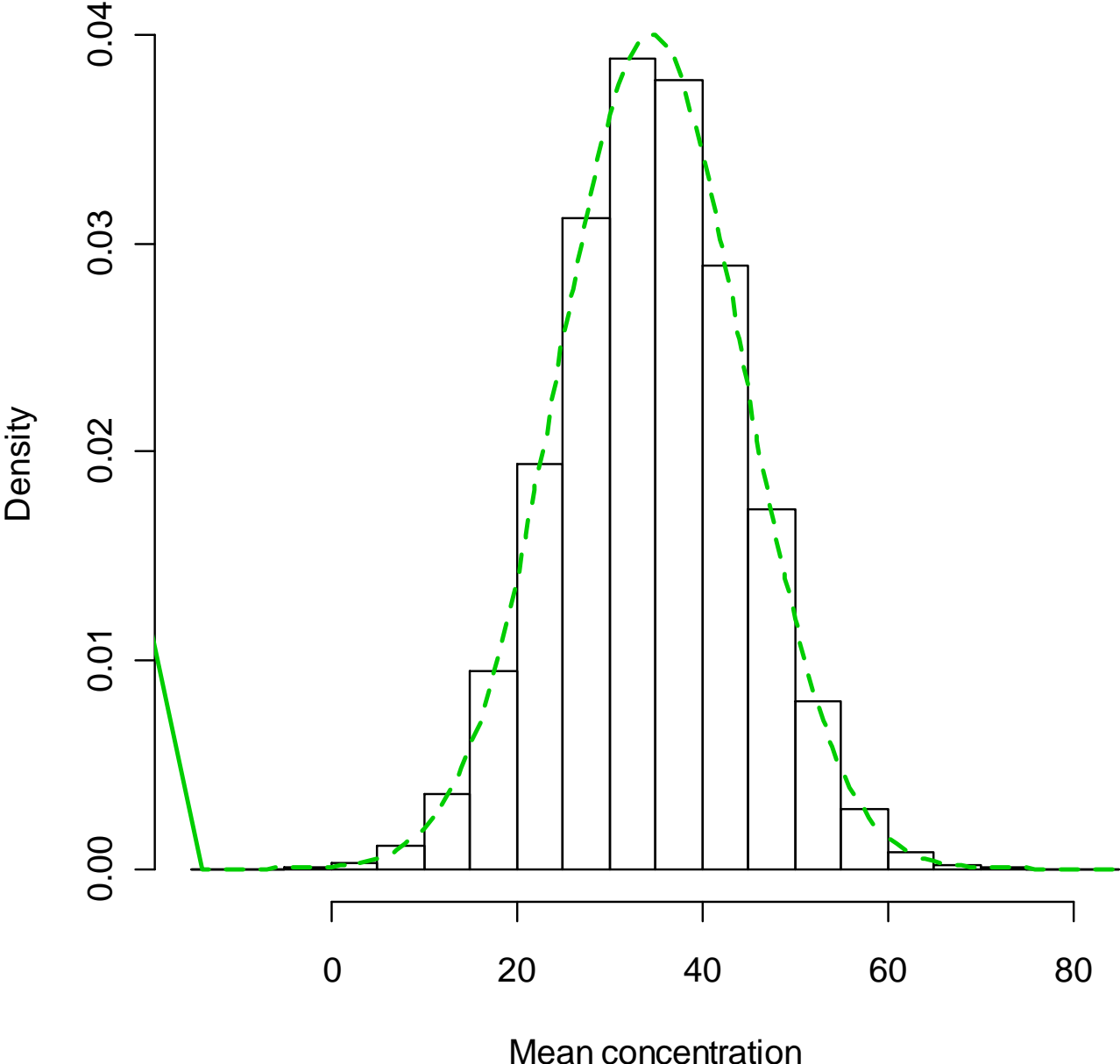
Results

- Mean= $(32.9 + 35.9 + 35.0)/3 = 34.5$
- SE=SD(32.9, 35.5, . . . , 35.0) = 9.97
- Mean concentration 34.5 ± 10
(mean \pm SE)
 - What does this mean?

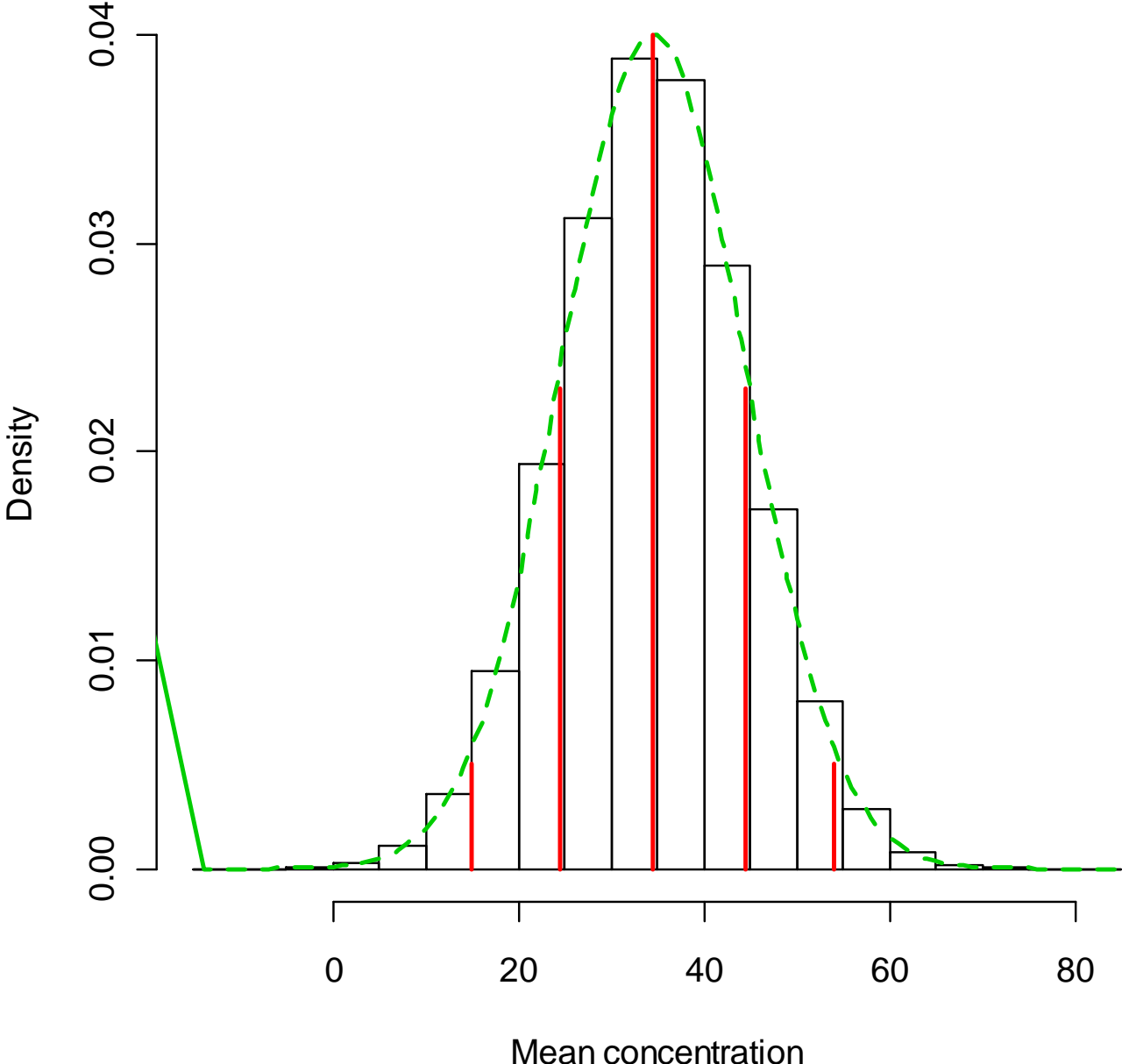
Distribution of Means of Salt Concentrations



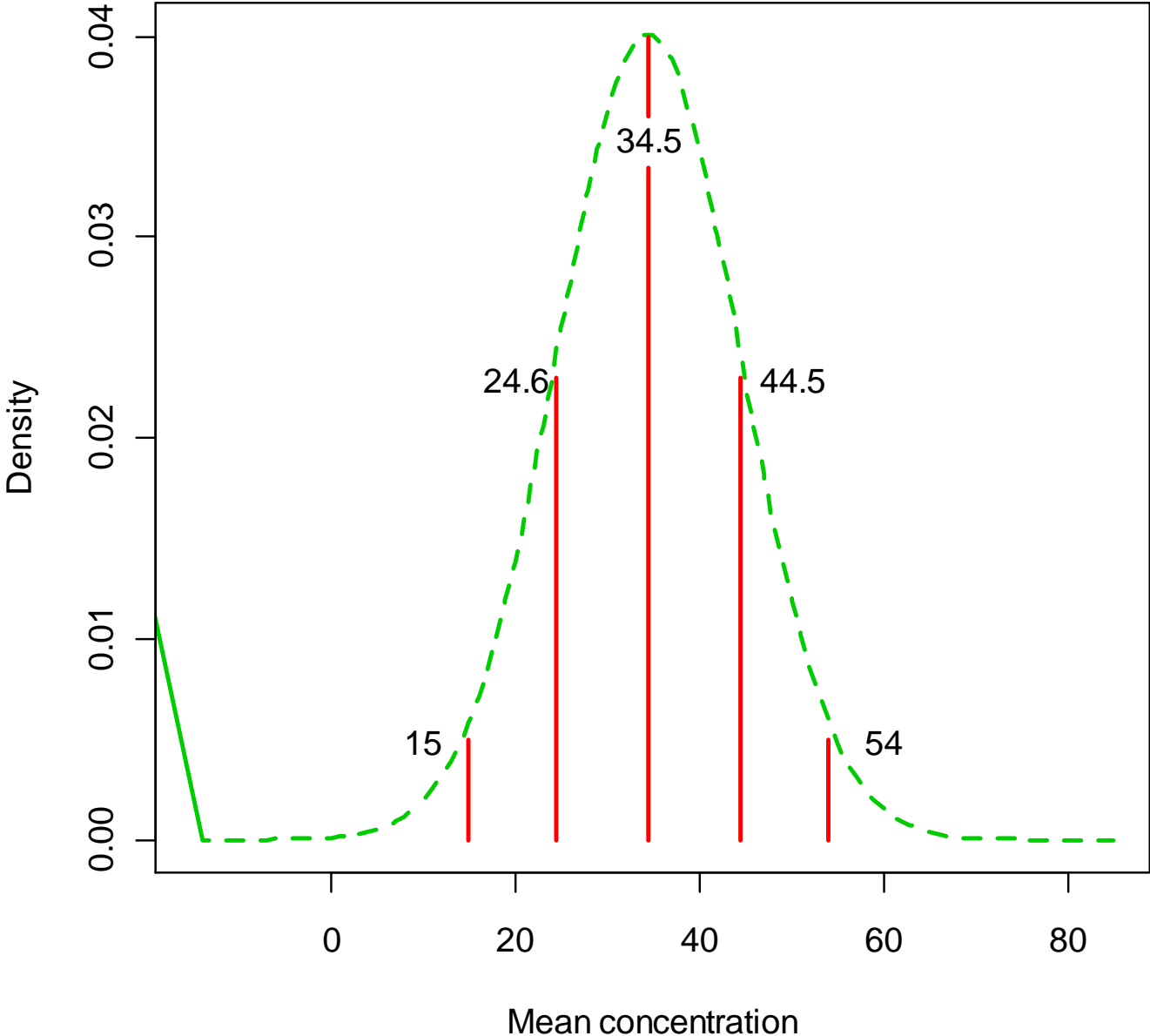
Distribution of Means of Salt Concentrations



Distribution of Means of Salt Concentrations



Distribution of Means of Salt Concentrations



What is the salinity of Ocean Water?

Standard Error

Relation to Standard Deviation

- A **histogram** of the **means** will follow a **normal distribution**
 - (Central Limit Theorem)
- The standard deviation of the means is the standard error of the mean
 - $SE = SD(\text{mean}_1, \text{mean}_2, \dots, \text{mean}_n)$
 - In our example, the standard deviations of the **mean concentrations** from Baltimore, Tahiti, Hawaii, etc.

Standard Error

- The standard error of the sample means can be **estimated** from the **SD of the observations in a single sample**

$$SE = SD / \sqrt{n}$$

- Where n is the number of observations in the sample

Standard Error

- In our case, $SE = SD/\sqrt{10}$
- We could have measured sodium in a 10 samples from a single location
 - (not have traveled the world)
- Computed a mean
- Computed a standard deviation
- Have an estimate of the standard error.

$$SE = SD/\sqrt{10}$$

Standard Error

- When is the standard error used?
 - When there is a desire to **compare means**
 - The SE is the variability of distribution of sample means

Standard Error

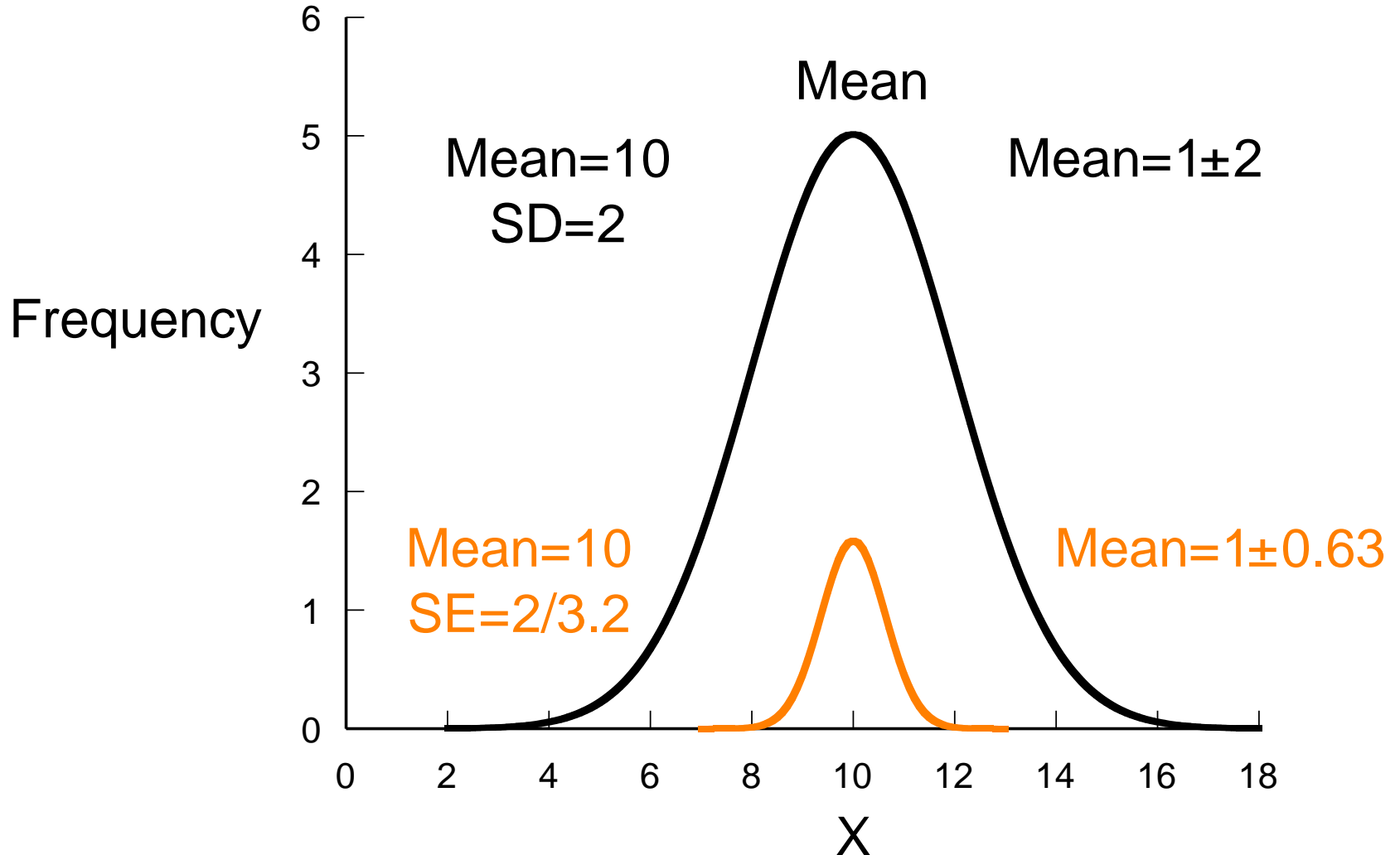
- When is the standard error used?
 - When there is a desire to **compare means**
 - The SE is the variability of distribution of sample means

$$t = \frac{\text{mean}_2 - \text{mean}_1}{SE_{\text{Difference}}} \qquad t = \frac{80 - 40}{32} = 1.3$$

$$p(t_{1.3,38}) = 0.21$$

Standard Error

From a Sample of 10 Measurements



SD vs SE

- Mean and SD
 - Describes distribution of your 10 measurements
- Mean and SE
 - Describes the distribution of a series of means obtained by random draws from your 10 observations.

Standard Deviation vs. Standard Error

- Use
 - Standard deviation;
 - To describe the **variability** of a characteristic **in a group of subjects.**
 - Standard error
 - To describe the **variability** of the mean of a characteristic
 - When **comparing groups** of subjects. (It describes the precision of each group's mean.)

Standard Deviation vs. Standard Error

- We recruited 100 subjects, 50 men and 50 women, mean age 50 ± 10 (mean \pm SD) yr. On average the men were taller than the women, 173 ± 0.71 vs. 157 ± 0.64 cm (mean \pm SE), $p < 0.0001$.
- 50 ± 10 distribution of subject's age
- 173 ± 0.71 , 157 ± 0.64 distribution of the mean of height

What is the salinity of Ocean Water?

- Mean 35 ppt (parts/thousand), range 33 to 37 ppt
 - Or
- 35 ± 0.56 ppt (mean \pm SE)



Visualization of Data

Visualization of Data

- Location
- Shape
- Range
- Extreme Values

Visualization of Data

- Location
 - Where are most of the data?
- Shape
 - What form does the data take?
- Range
 - What are the largest and smallest values?
- Extreme Values
 - Are there values are extremely large or small?

Visualization of Data

- Stem and leaf diagram
- Histogram
- Density plot
- Box plot
- QQ plot

Visualization of Data

- For a small batch we can list the data:
 - 1, 2, 3, 4, 5, 6, 7, 8, 9, 10
- Summarization: With a measure of centrality and spread
 - Mean \pm SD = 5.5 \pm 3.0

List fail for a Big Batch

- 48.70784 51.29260 48.54744 50.70652 47.73479 50.95584 51.40754
- 48.80511 50.15987 48.13988 50.50101 49.12918 50.79219 48.95935
- 49.47898 51.95629 50.69902 49.49839 49.09733 51.36002 49.86675
- 50.48088 50.72452 50.95122 49.59450 50.94099 49.86079 49.03003
- 50.91817 49.38528 47.99457 50.75871 48.81016 51.65646 50.67613
- 48.40501 49.55248 49.74465 50.98077 48.65700 48.84910 51.40241
- 49.15870 49.76532 51.14845 47.88810 50.48360 49.70937 49.25858
- 47.71131 49.73282 48.68164 51.88570 49.03893 49.76663 49.36683
- 49.47986 50.16270 49.17999 49.02746 50.80737 50.54097 49.55330
- 49.26511 49.88770 51.50638 51.16936 48.49381 50.05837 51.39448
- 50.03687 50.55491 50.80347 52.99833 50.73491 50.22002 51.22859
- 49.75077 49.87826 48.30033 51.54809 49.09882 50.37503 50.29963
- 48.80889 50.09534 51.42604 49.75496 51.61375 49.77668 51.73419
- 49.61100 49.92117 49.98774 48.71677 51.47969 50.86786 50.63124
- 50.17466 50.94377

Summary of Batch

- $N=100$
- $\text{Mean} \pm \text{SD} = 50 \pm 0.85$

Summary of Batch

- $N=100$
- $\text{Mean} \pm \text{SD} = 50 \pm 0.85$
- $\text{Range} = 48 \text{ to } 52$

Stem and Leaf Diagram

Decimal point is at the colon

47 : 8

48 : 344

48 : 588899999

49 : 0111112222222333444

49 : 556666667777777788889

50 : 000011112222334444

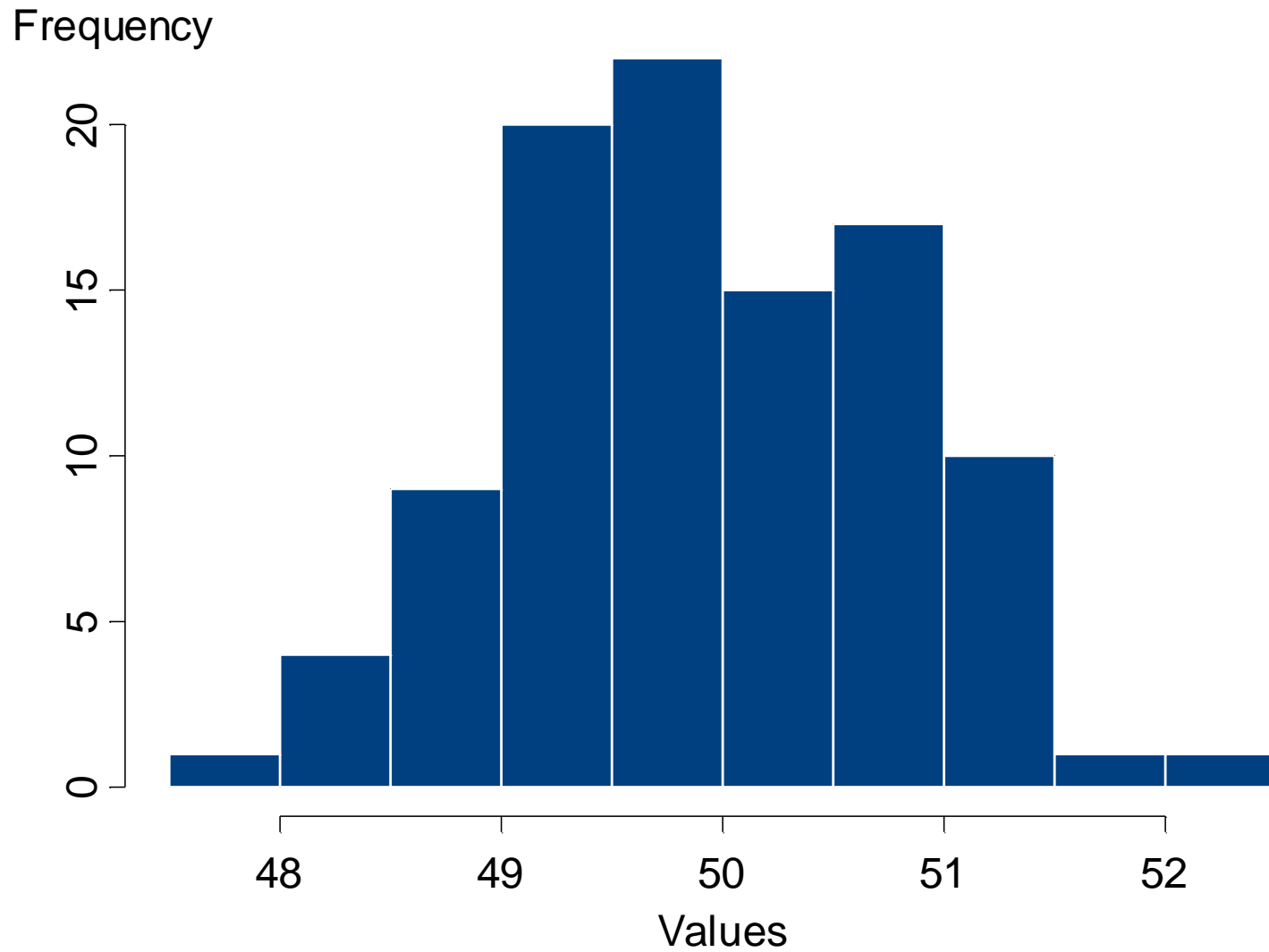
50 : 5566666667778889

51 : 01112223444

51 : 7

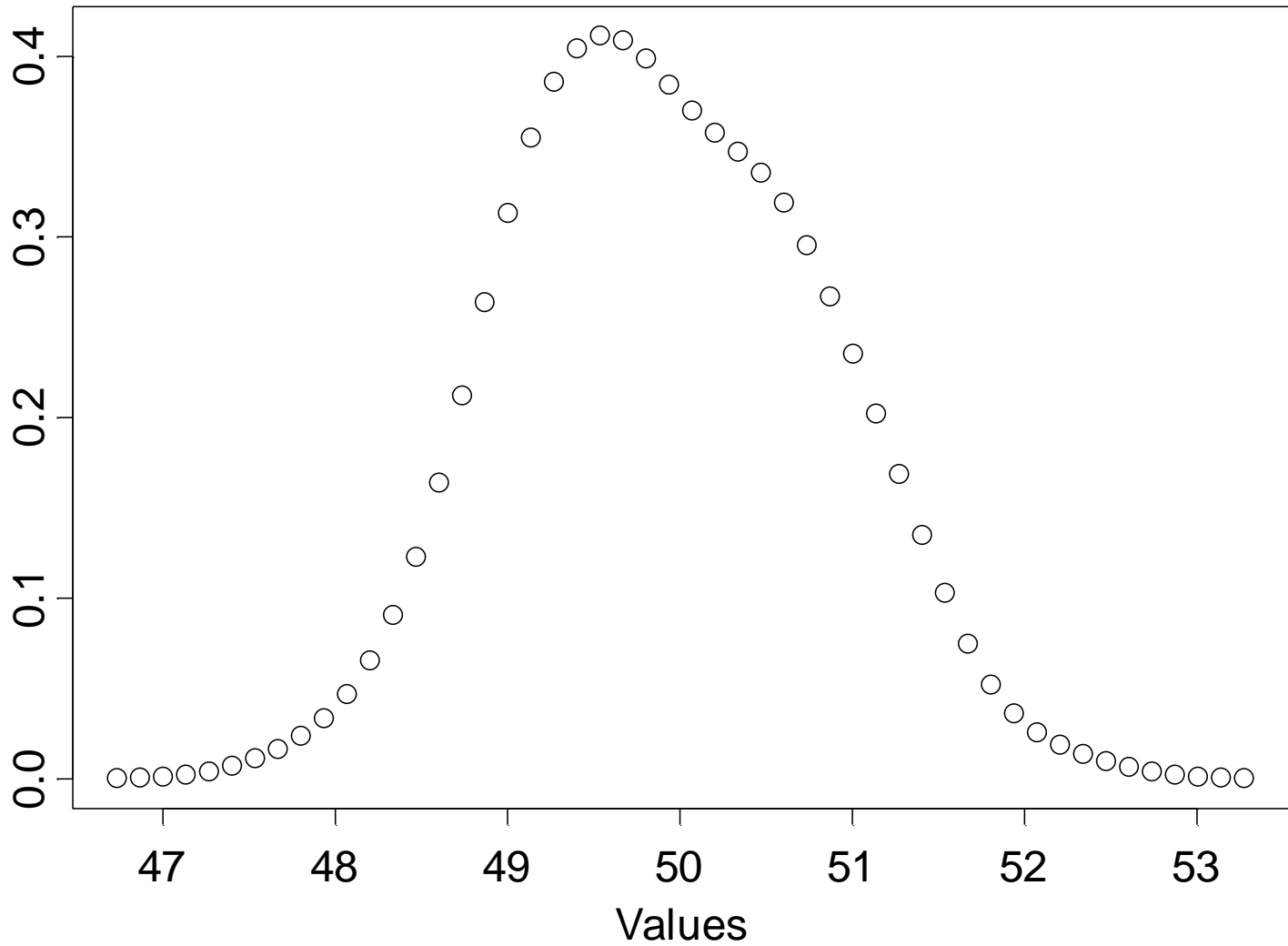
52 : 2

Histogram



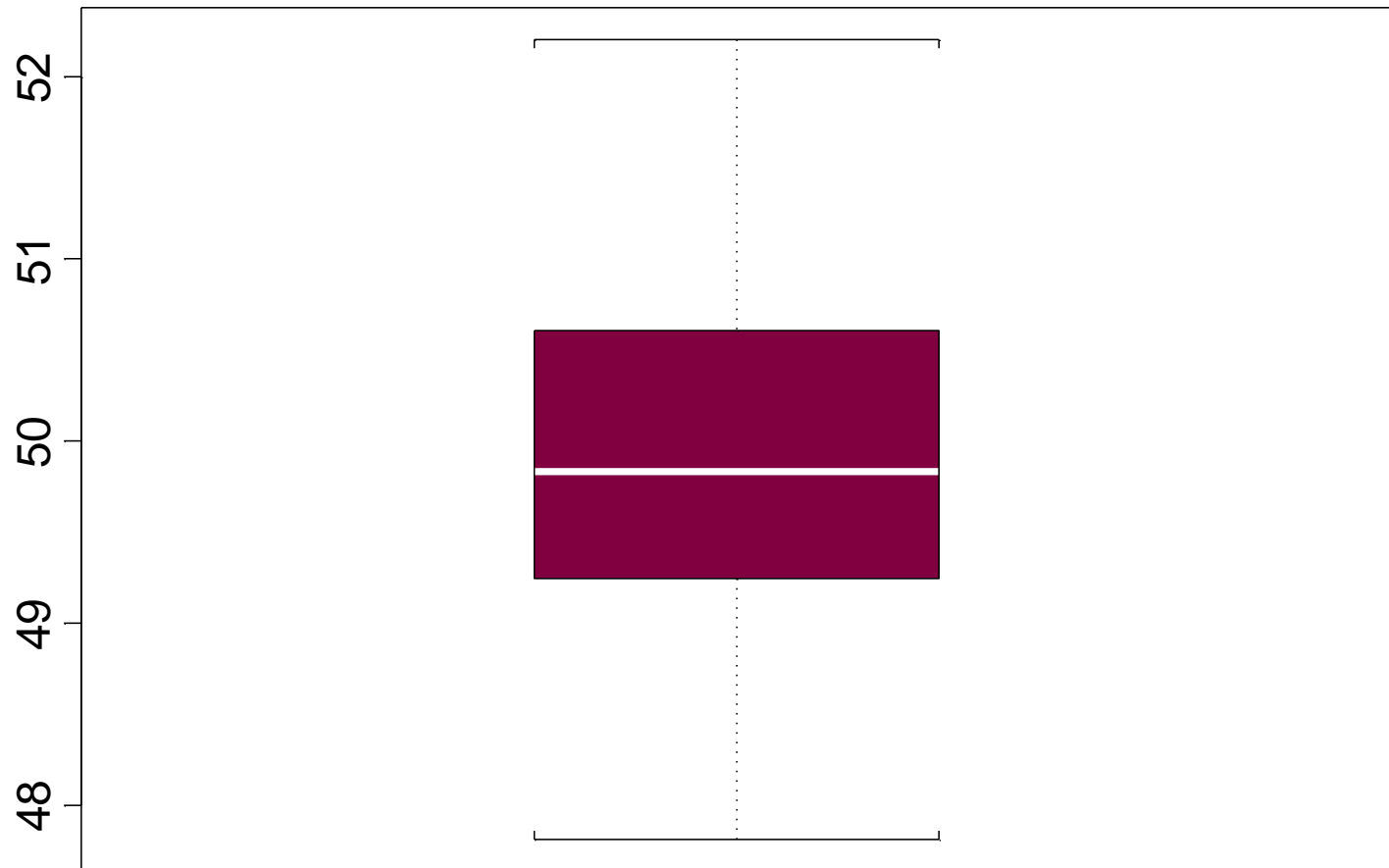
Density Plot

Probability



Box Plot

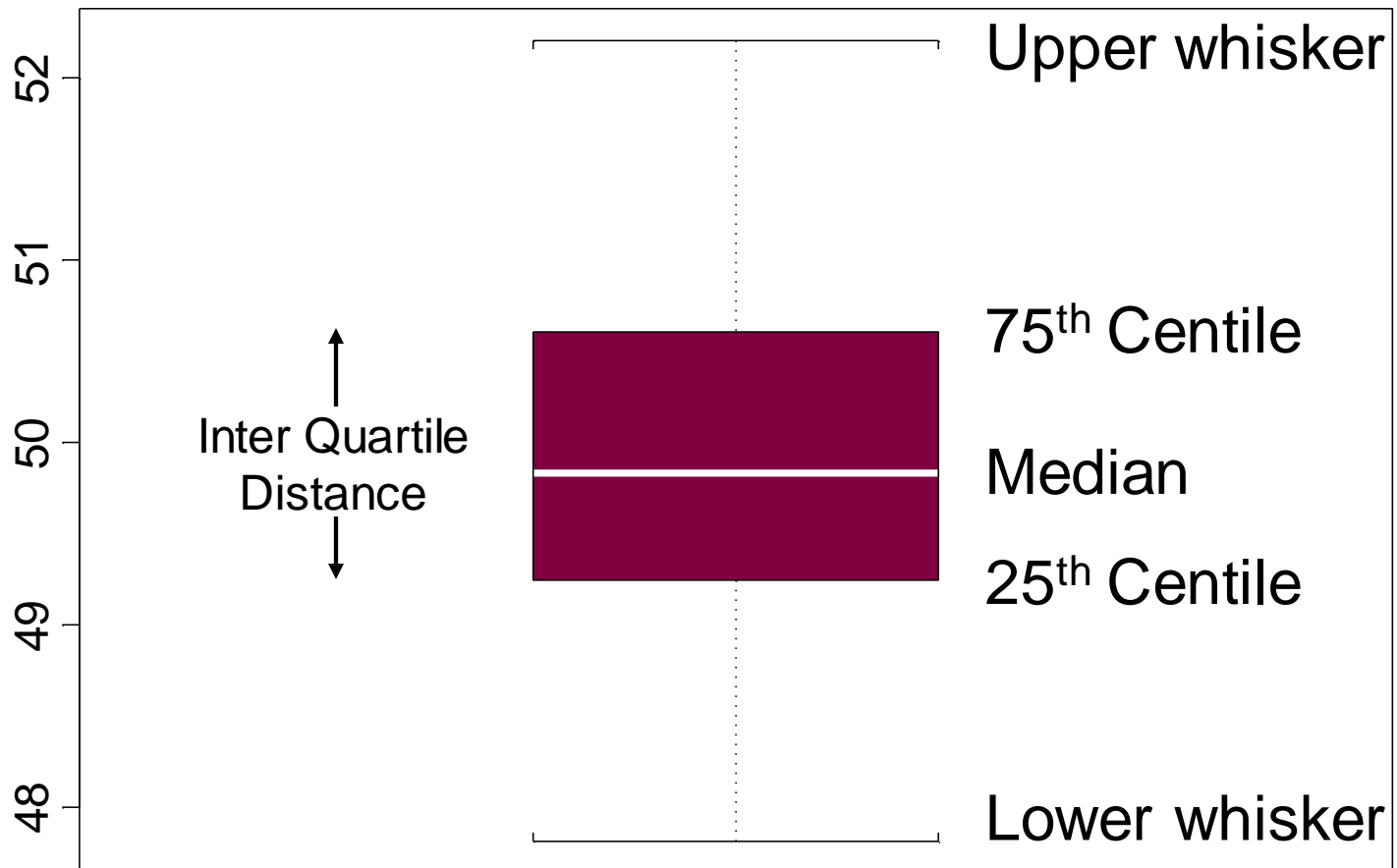
Values



Box Plot

(Box and Whisker Plot)

Values



John W. Tukey

1915-2000



- American statistician
 - Home schooled
 - Brown University
 - Masters Chemistry
 - Princeton
 - Ph.D. Mathematics
 - Faculty Member
 - Bell Labs
- Invented
 - Boxplot
 - Stem and leaf diagram
- Introduced in his book,
 - *Exploratory Data Analysis* (1977)

Inter Quartile Distance (IQD)

- IQD (Inter quartile distance)
 - $\text{IQD} = 4^{\text{th}} \text{ quartile} - 1^{\text{st}} \text{ quartile}$
 - or
 - $\text{IQD} = 75^{\text{th}} \text{ centile} - 25^{\text{th}} \text{ centile}$

Number	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Centile	6	12	18	24	29	35	41	47	53	59	65	71	76	82	88	94	100

- $\text{IQD} = 12 - 4 = 8$

The IQD is Resistant to Extreme Values

Number	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Centile	6	12	18	24	29	35	41	47	53	59	65	71	76	82	88	94	100

- $IQD = 12 - 4 = 8$

Number	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	1600
Centile	6	12	18	24	29	35	41	47	53	59	65	71	76	82	88	94	100

- $IQD = 12 - 4 = 8$

Resistant Statistic

- A statistic that is
 - relatively unchanged
 - when a large change is made
- in a small fraction of the data that are used to compute the statistic

The Mean and Standard Deviation are Not Resistant

Number	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Centile	6	12	18	24	29	35	41	47	53	59	65	71	76	82	88	94	100

- Mean=8, SD=5

Number	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	1600
Centile	6	12	18	24	29	35	41	47	53	59	65	71	76	82	88	94	100

- Mean=101, SD=386

IQD Used to Create a Resistant Standard Deviation

- IQD=75th centile-25th centile

$$\text{IQD} = \mu + 0.67\sigma - (\mu - 0.67\sigma)$$

$$\text{IQD} = 1.35\sigma$$

- Thus

$$\sigma = \text{IQD}/1.35$$

- Or for an IQD of 8,

$$\sigma = 8/1.35 = 5.9$$

The Median and IQD Standard Deviation are Resistant

Number	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Centile	6	12	18	24	29	35	41	47	53	59	65	71	76	82	88	94	100

- Mean=8, SD=5
- IQR=8
- Median=7.5 IQDSD=5.9

Number	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	1600
Centile	6	12	18	24	29	35	41	47	53	59	65	71	76	82	88	94	100

- Mean=101, SD=386
- IQR=8
- Median=7.5 IQDSD=5.9

Box Plot

Fences Based on IQD SD

Upper Fence = 75th centile + 1.5 * IQD

$$\mu + 0.67\sigma + 1.5 * 1.35\sigma$$

$$\mu + 2.02\sigma$$

Lower Fence = 25th centile - 1.5 * IQD

$$\mu - 0.67\sigma - 1.5 * 1.35\sigma$$

$$\mu - 2.02\sigma$$

- Mean \pm 2 SE includes ~98% of the data
- Fences **enclose** ~98% of the data

Box Plot

Fences Based on IQD SD

Upper Fence (UF)=75th centile + 1.5*IQD

$$\mu + 0.67\sigma + 1.5 * 1.35\sigma$$

$$\mu + 2.02\sigma$$

Draw whisker at largest value smaller than UF

Lower Fence (LF)=25th centile - 1.5*IQD

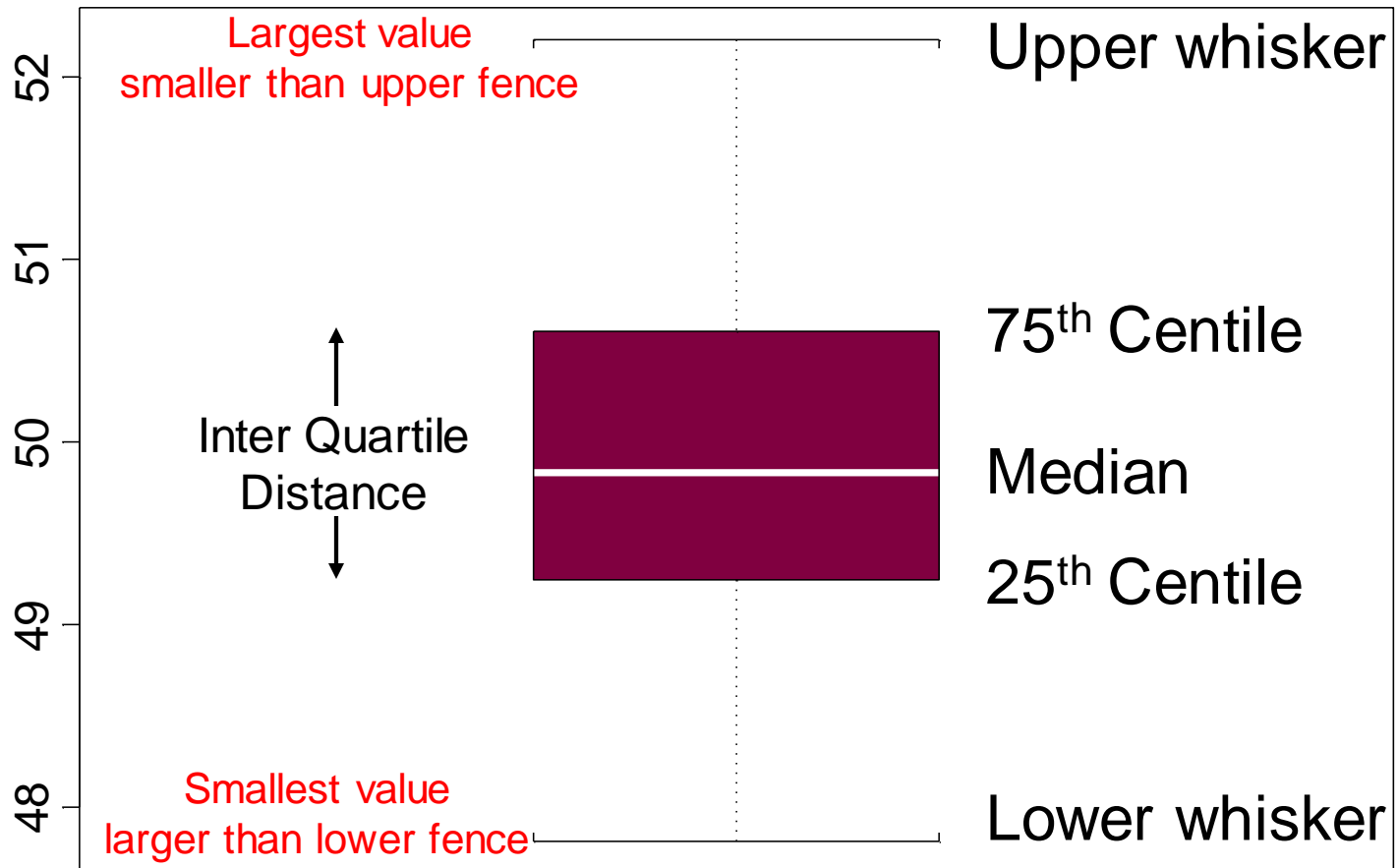
$$\mu - 0.67\sigma - 1.5 * 1.35\sigma$$

$$\mu - 2.02\sigma$$

Draw whisker at smallest value larger than LF

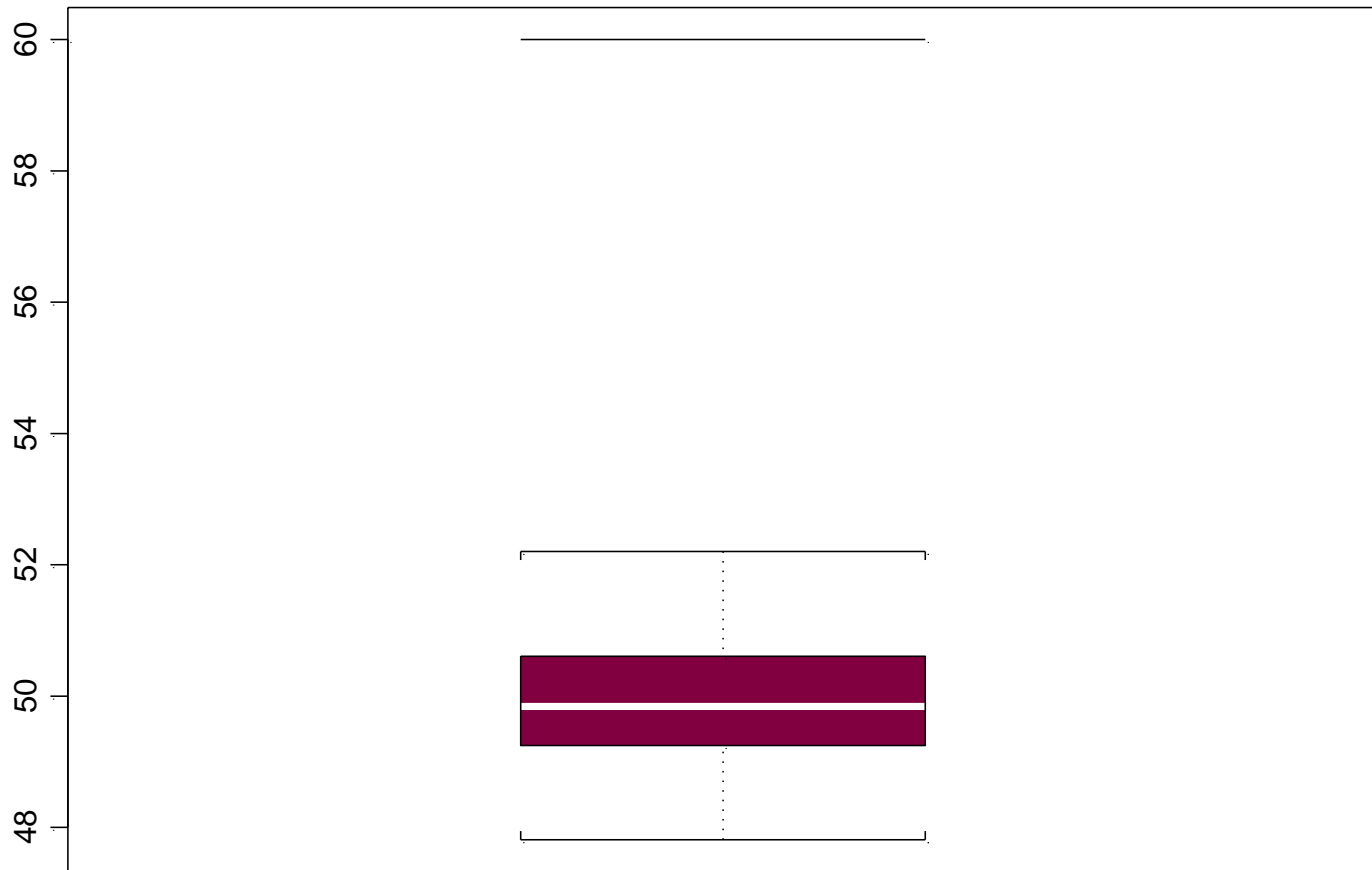
Box Plot

Values



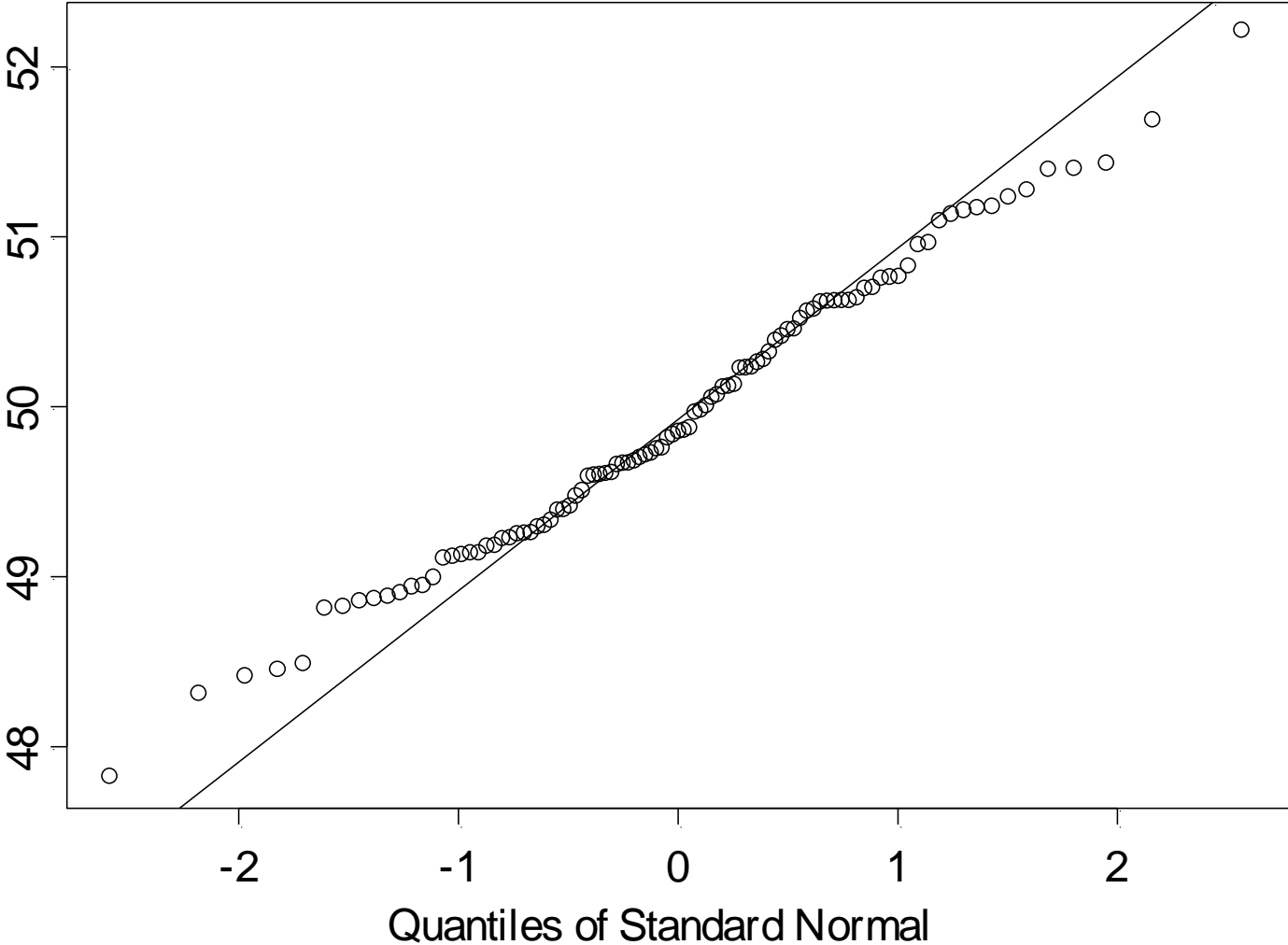
Box Plot with Extreme Value

Value

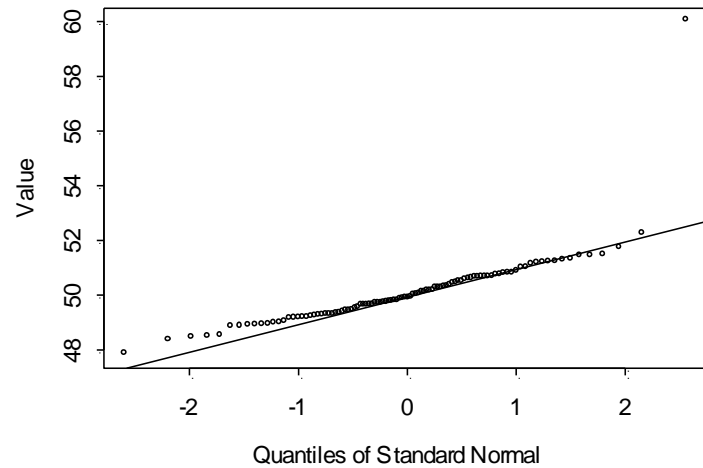
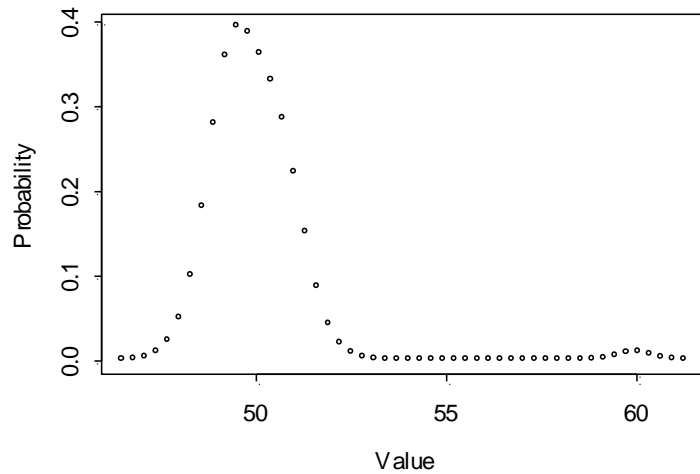
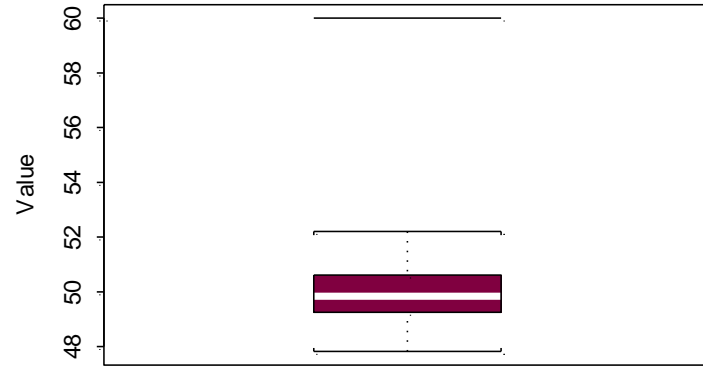
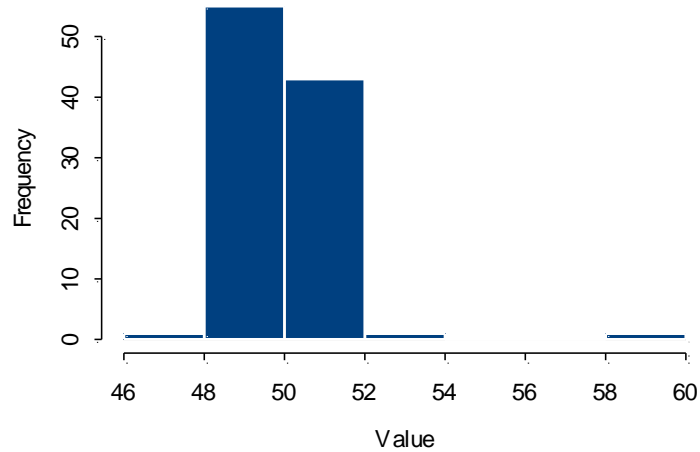


QQ plot

Value



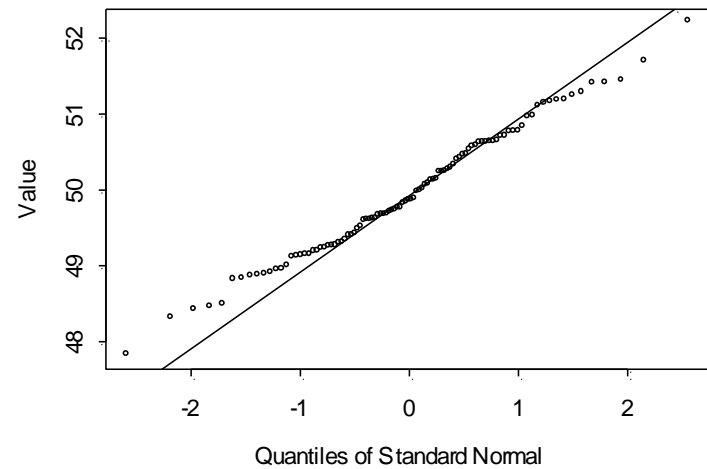
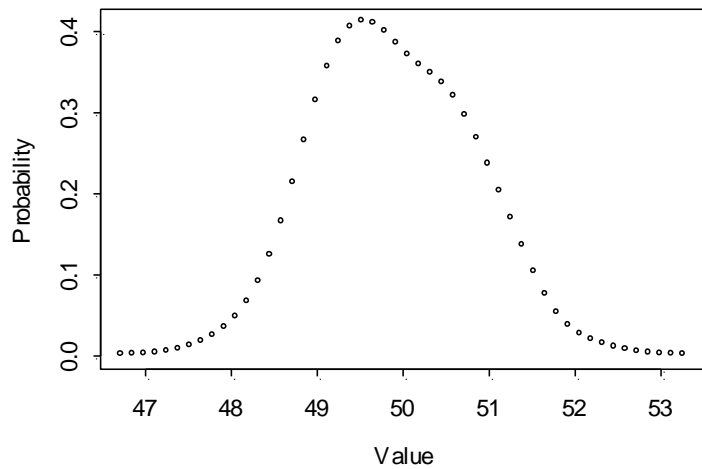
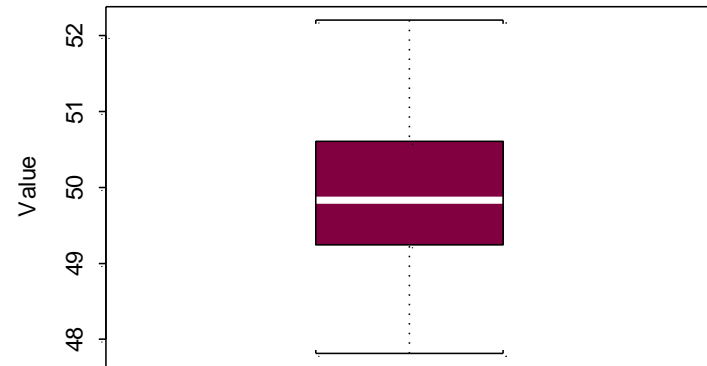
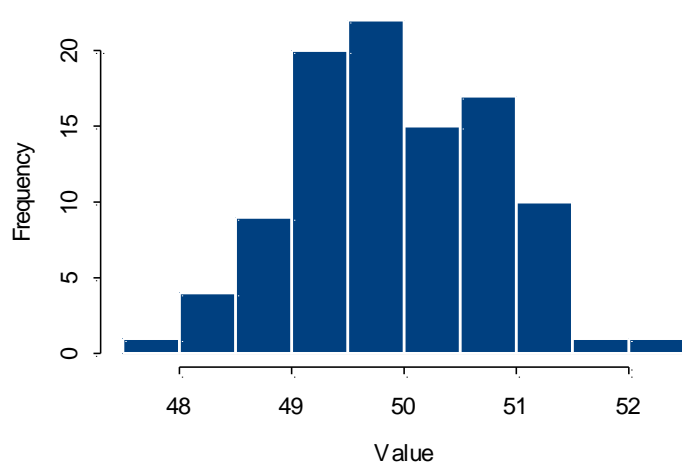
Batch with Extreme Value



Visualization of a Batch

- 48.70784 51.29260 48.54744 50.70652 47.73479 50.95584 51.40754
- 48.80511 50.15987 48.13988 50.50101 49.12918 50.79219 48.95935
- 49.47898 51.95629 50.69902 49.49839 49.09733 51.36002 49.86675
- 50.48088 50.72452 50.95122 49.59450 50.94099 49.86079 49.03003
- 50.91817 49.38528 47.99457 50.75871 48.81016 51.65646 50.67613
- 48.40501 49.55248 49.74465 50.98077 48.65700 48.84910 51.40241
- 49.15870 49.76532 51.14845 47.88810 50.48360 49.70937 49.25858
- 47.71131 49.73282 48.68164 51.88570 49.03893 49.76663 49.36683
- 49.47986 50.16270 49.17999 49.02746 50.80737 50.54097 49.55330
- 49.26511 49.88770 51.50638 51.16936 48.49381 50.05837 51.39448
- 50.03687 50.55491 50.80347 52.99833 50.73491 50.22002 51.22859
- 49.75077 49.87826 48.30033 51.54809 49.09882 50.37503 50.29963
- 48.80889 50.09534 51.42604 49.75496 51.61375 49.77668 51.73419
- 49.61100 49.92117 49.98774 48.71677 51.47969 50.86786 50.63124
- 50.17466 50.94377

Visualization of a Batch



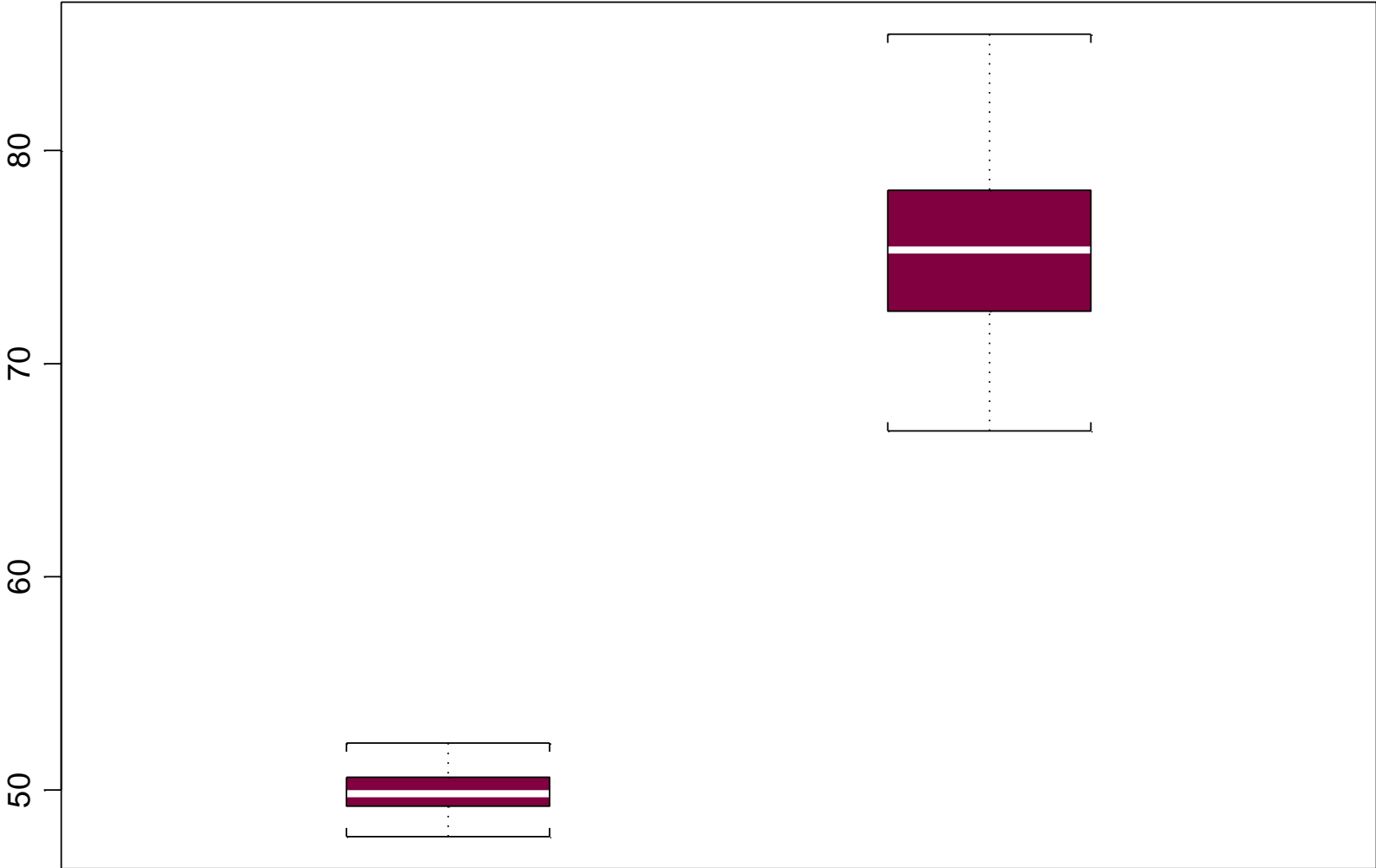
Comparison of Two Batches

- Batch 1, N=100
- Mean \pm SD = 50 \pm 0.85
- Range = 48 to 52

- Batch 2, N=100
- Mean \pm SD = 75 \pm 3.6
- Range = 67 to 85

Comparison of Two Batches

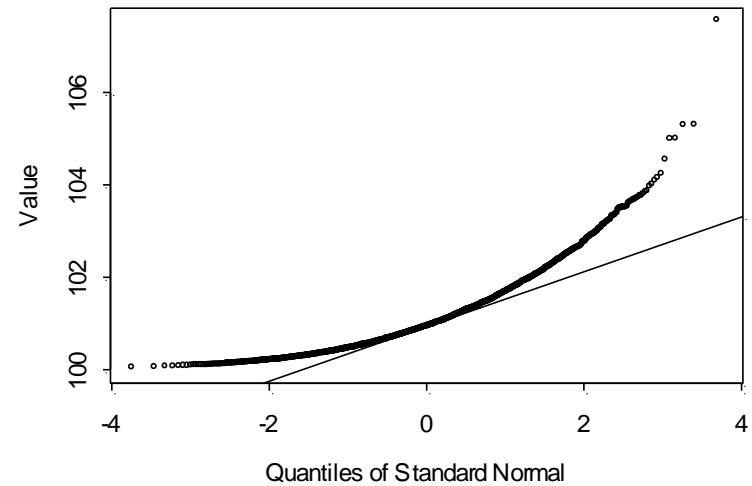
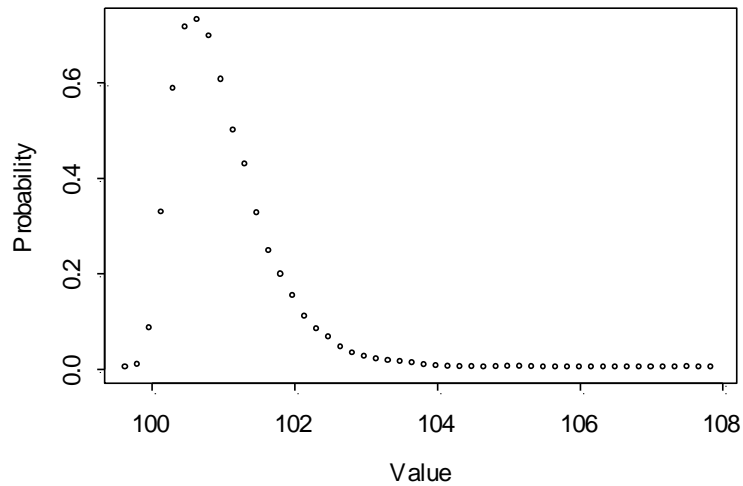
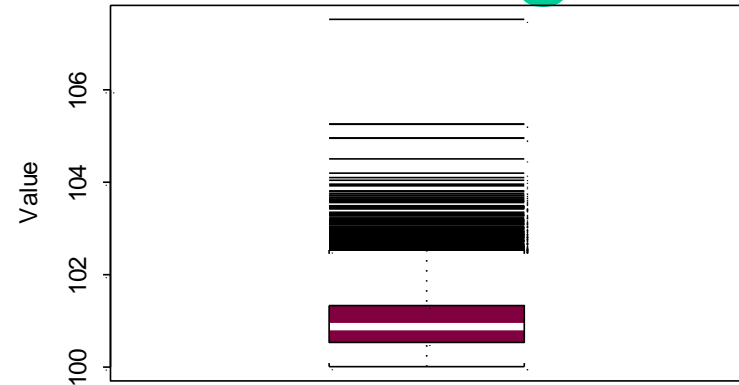
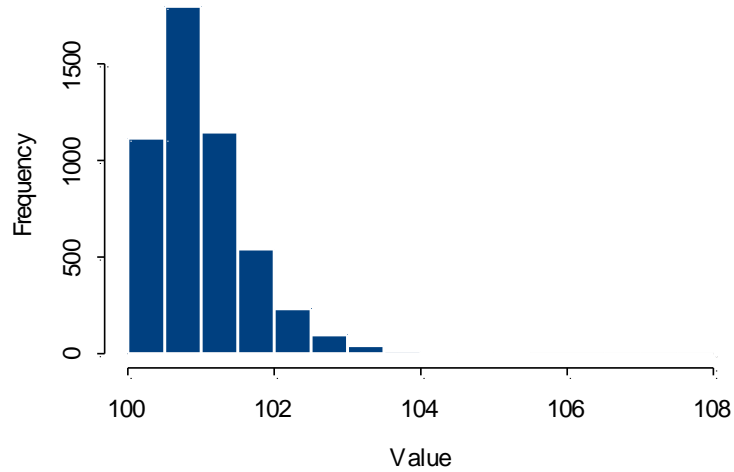
Value



Batch Skewed to the Right (High Values)

- $N=1000$
- $\text{Mean} \pm \text{SD} = 101 \pm 0.66$
- $\text{Range} = 100 \text{ to } 107.5$
- **Skew: Not symmetrical about the mean**

Batch Skewed to the Right



Resistance of SD based on IQR

- Batch 1
- 49 52 50 50 50 50 49 51 49 50 49 52 48
- 49 50 50 52 49 51 51 52 50 50 49 50

- Batch 2
- 49 52 50 50 50 50 49 51 49 50 49 52 48
- 49 50 50 52 49 51 51 52 50 50 49 **70**

Resistance of SD based on IQR

- Batch 1, N=25
- Mean \pm SD = 50 \pm 0.95

- Batch 2, N=25
- Mean \pm SD = 51 \pm 4.0

Resistance of SD based on IQR

- Batch 1, N=25
- Mean \pm SD = 50 \pm 0.95
- Range = 48 to 52

- Batch 2, N=25
- Mean \pm SD = 51 \pm 4.0
- Range = 48 to 70

Resistance of SD based on IQR

- Batch 1, N=25
- Mean \pm pSD = 50 \pm 0.87
- Batch 2, N=25
- Mean \pm pSD = 51 \pm 1.2

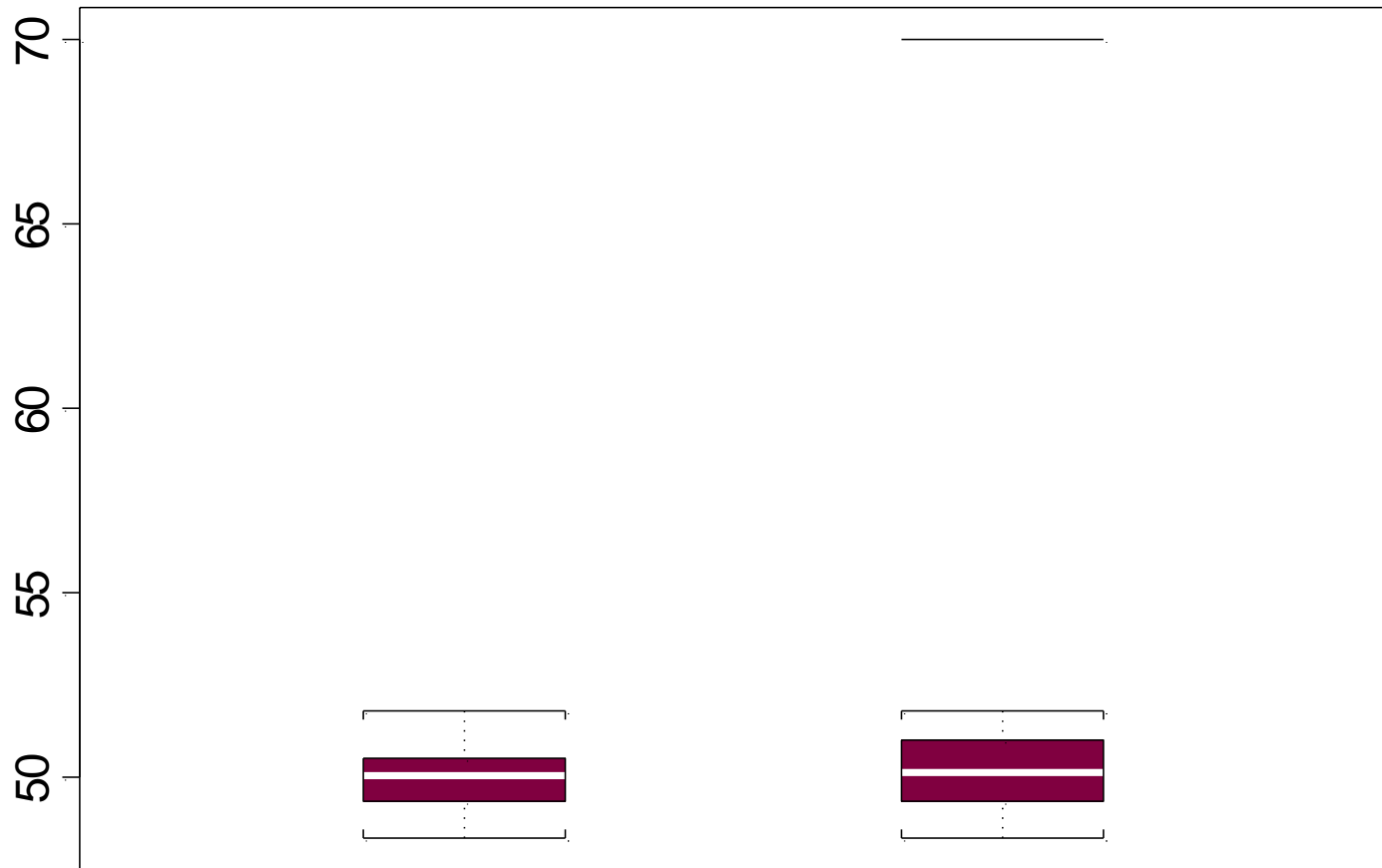
Resistance of SD based on IQR

- Batch 1, N=25
- Mean \pm SD = 50 \pm 0.95
- Mean \pm pSD = 50 \pm 0.87

- Batch 2, N=25
- Mean \pm SD = 51 \pm 4.0
- Mean \pm pSD = 51 \pm 1.2

Resistance of Box Plot

Value



A Comparison of Plots

	<u>Location</u>	<u>Shape</u>	<u>Range</u>	<u>Extreme Values</u>
Stem and leaf		x	x	x
Histogram		x	x	
Density		x		
Box	x		x	x
QQ		x	x	

Regression

Reducing Variance

A Batch of Data

40.60524	36.34975	43.19606	80.74827	81.01183
69.46780	95.49872	145.18474	111.62884	159.13814
147.96071	174.44359	262.73295	152.52425	192.91095
230.72459	198.48519	375.34976	362.29408	342.46495
400.57719	320.84187	240.24111	260.19998	276.23798
385.49059	355.32447	293.93133	351.88088	474.42359

Variance

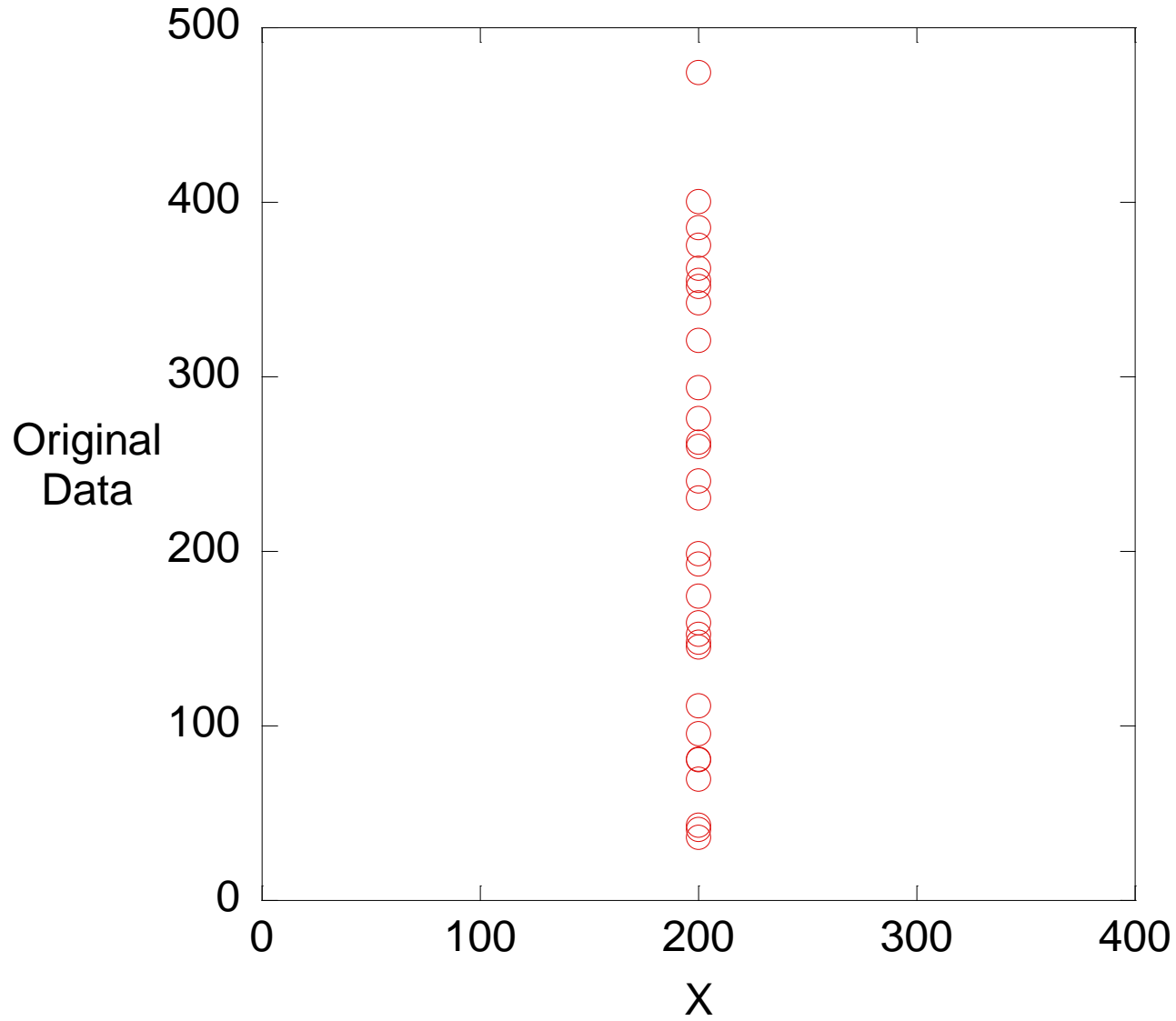
A Measure of Spread

Measures of Spread - Variance

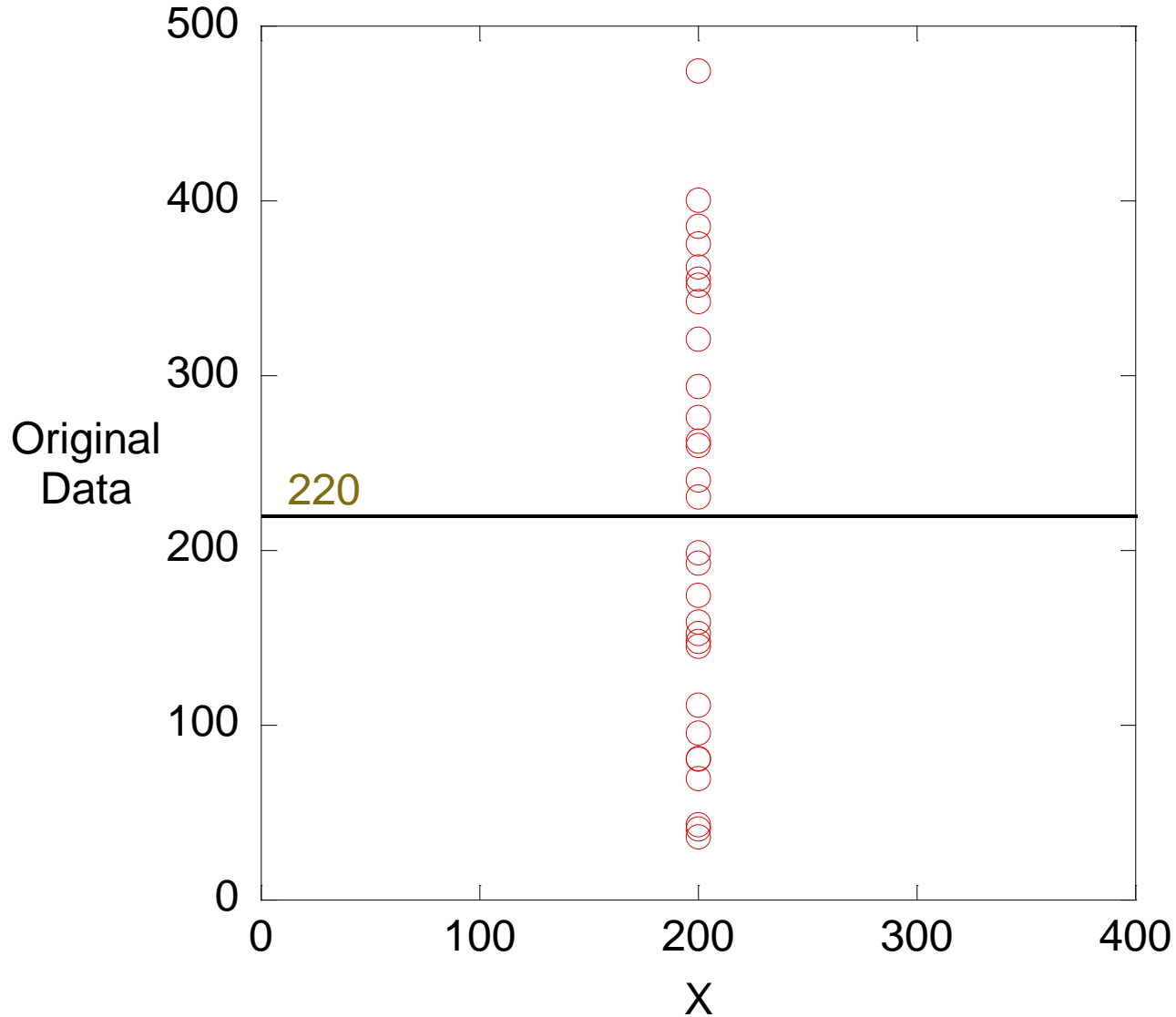
- Variance: The average of the squared distance from the mean

$$\sigma^2 = \sum_{i=1}^n (x_i - \bar{x})^2 / n \text{ or, } s^2 = \sum_{i=1}^n (x_i - \bar{x})^2 / (n - 1)$$

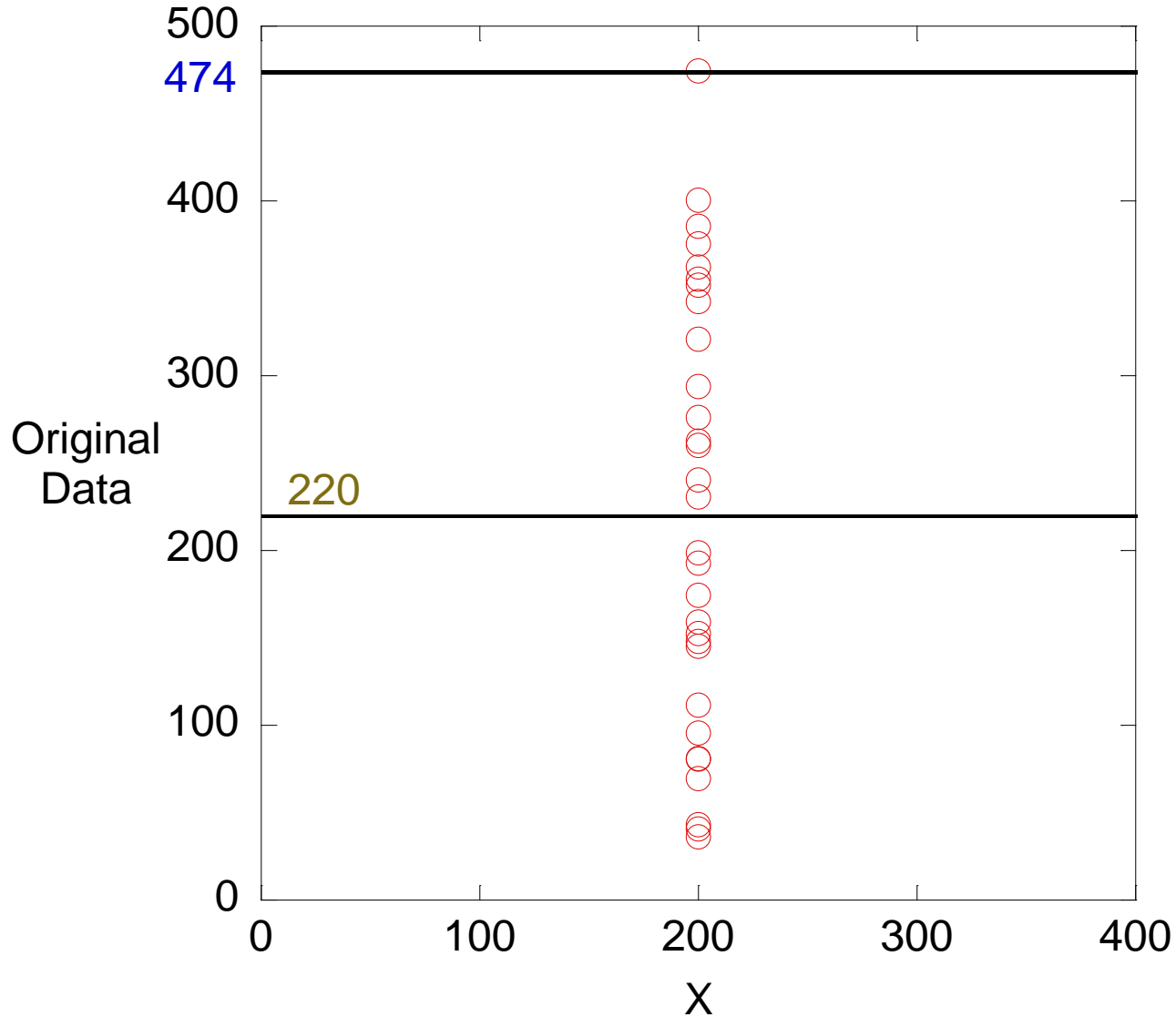
Original Data



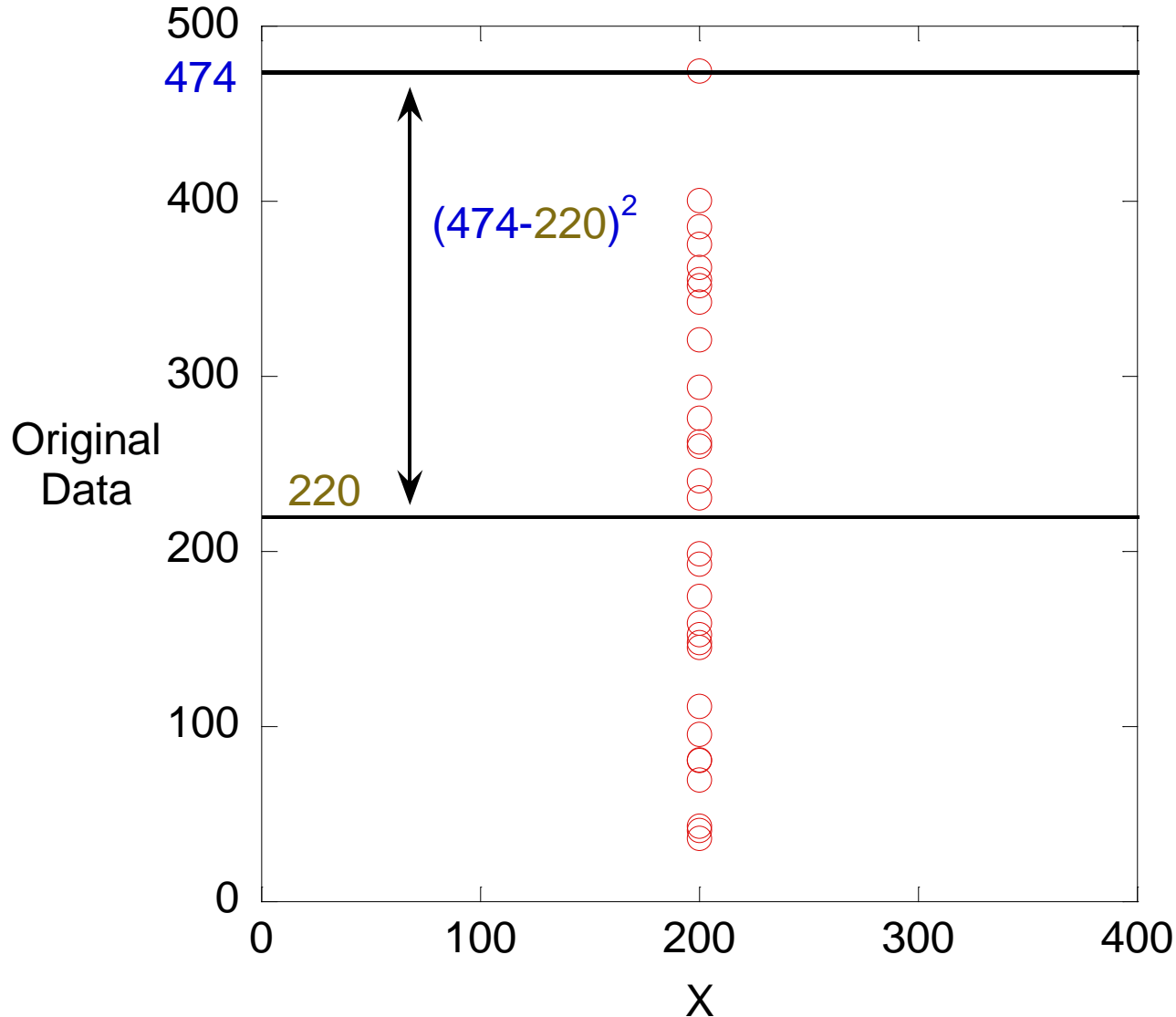
Mean Value



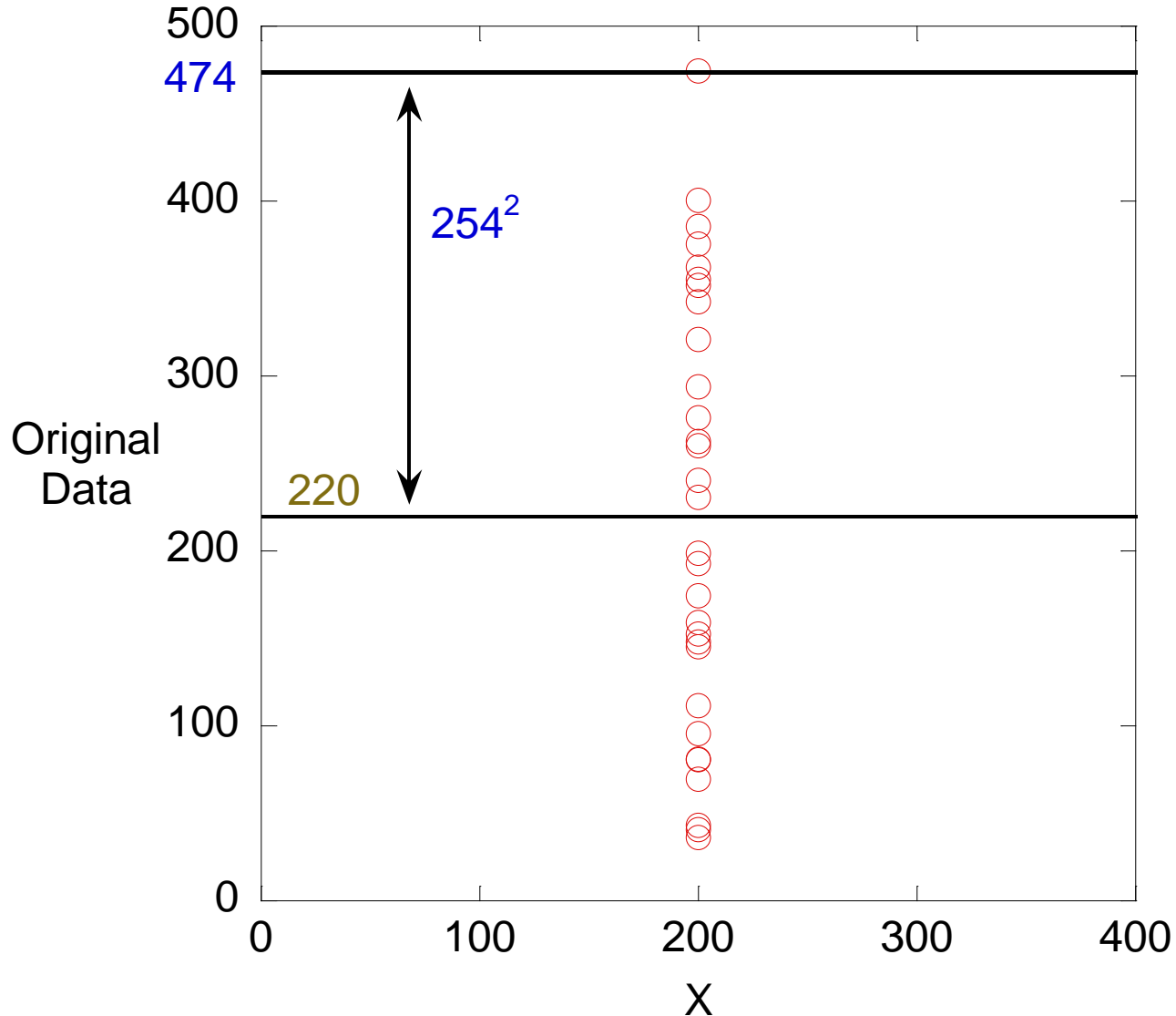
Largest Value



Squared Distance From Mean



Squared Distance From Mean



Variance

$$\sigma^2 = \sum_{i=1}^n (x_i - \bar{x})^2 / (n-1) = 15,536$$

Variance

$$s^2 = \sum_{i=1}^n (x_i - \bar{x})^2 / (n-1) = 15,536$$

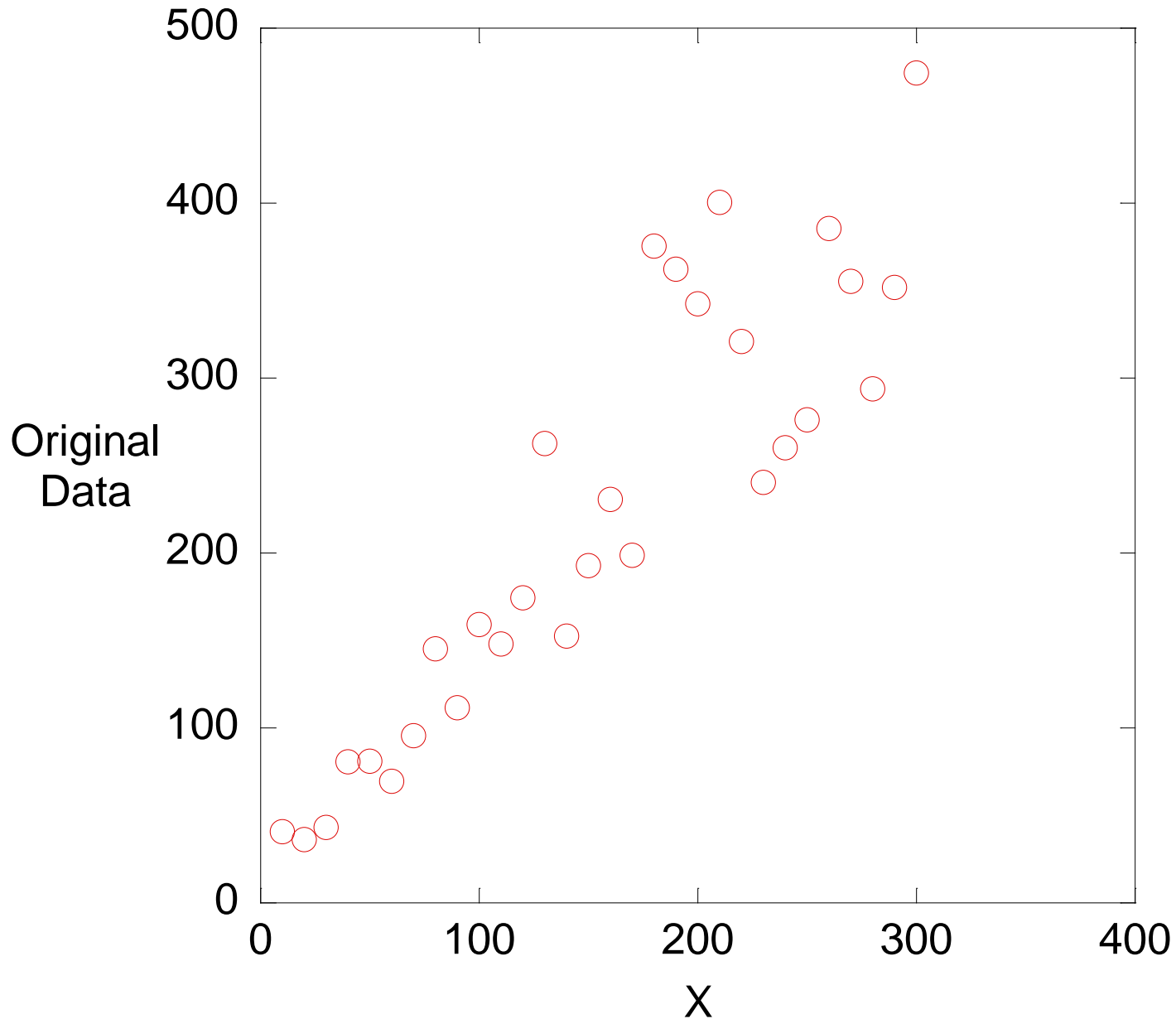
A Batch of Data

<u>y</u>	<u>y</u>	<u>y</u>
40.61	147.96	400.58
36.35	174.44	320.84
43.20	262.73	240.24
80.75	152.52	260.20
81.01	192.91	276.24
69.47	230.72	385.49
95.50	198.49	355.32
145.18	375.35	293.93
111.63	362.29	351.88
159.14	342.46	474.42

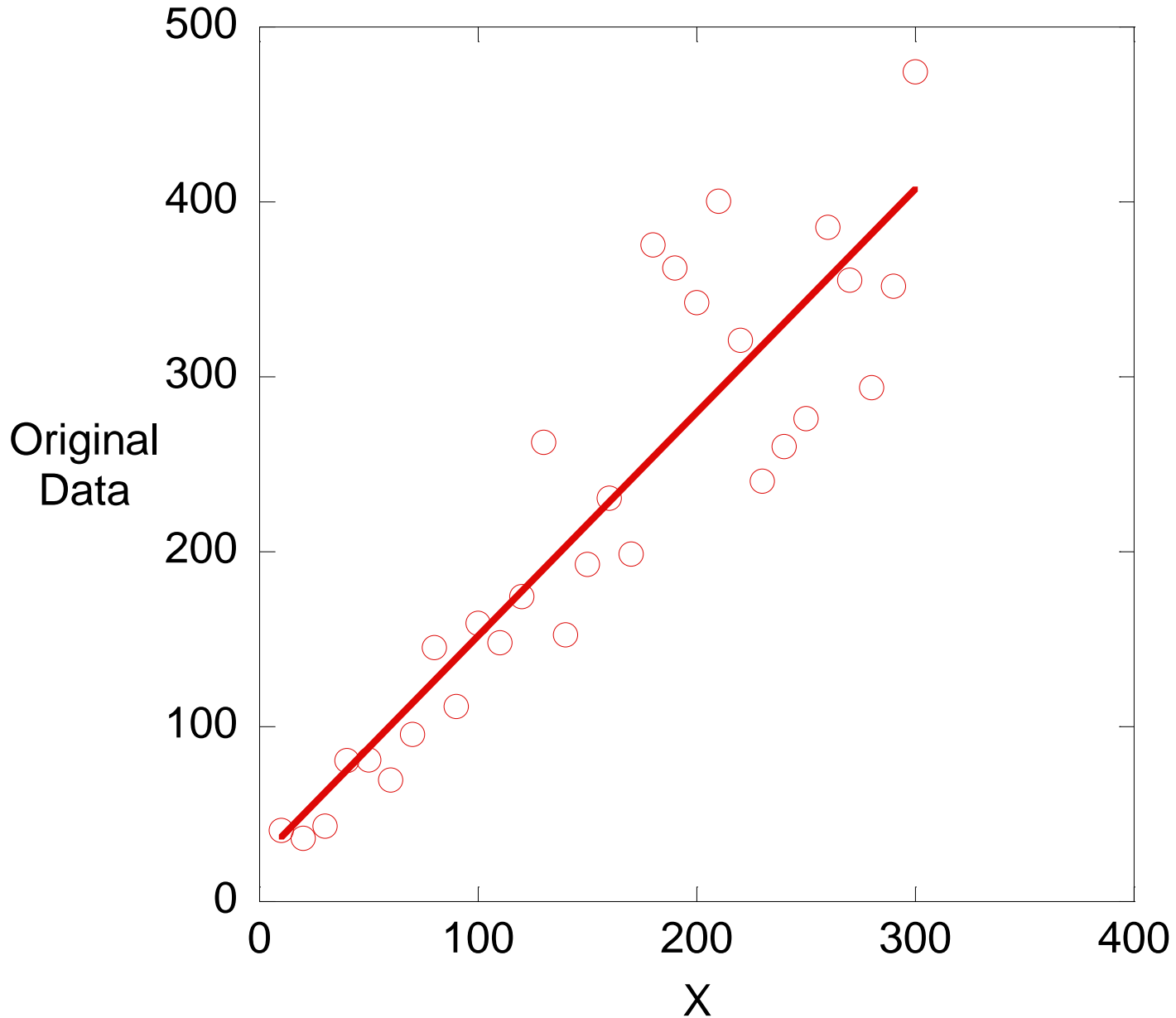
Each Y is Associated with An X

x	y	x	y	x	y
10	40.61	110	147.96	210	400.58
20	36.35	120	174.44	220	320.84
30	43.20	130	262.73	230	240.24
40	80.75	140	152.52	240	260.20
50	81.01	150	192.91	250	276.24
60	69.47	160	230.72	260	385.49
70	95.50	170	198.49	270	355.32
80	145.18	180	375.35	280	293.93
90	111.63	190	362.29	290	351.88
100	159.14	200	342.46	300	474.42

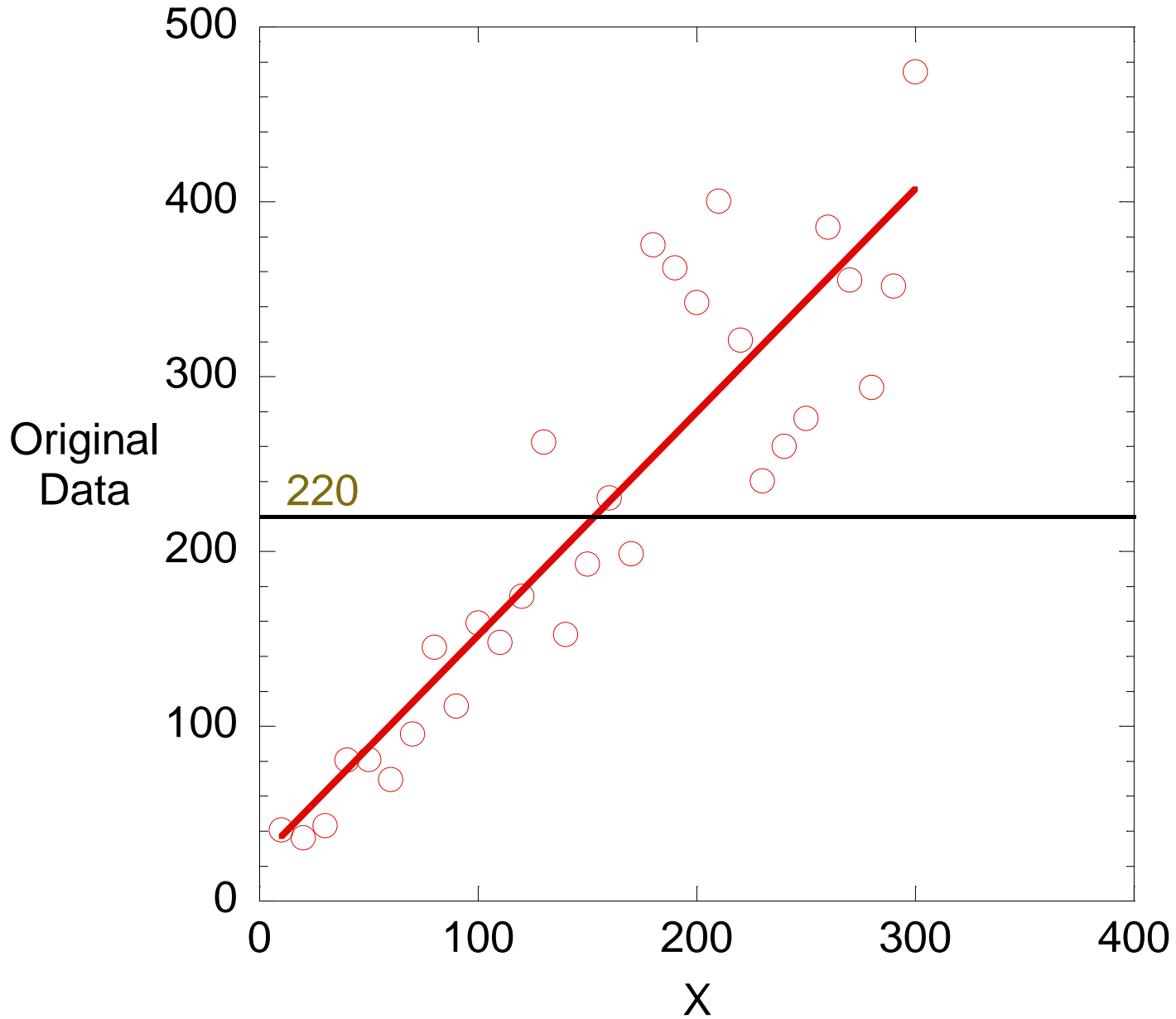
Each Y is Associated with a X



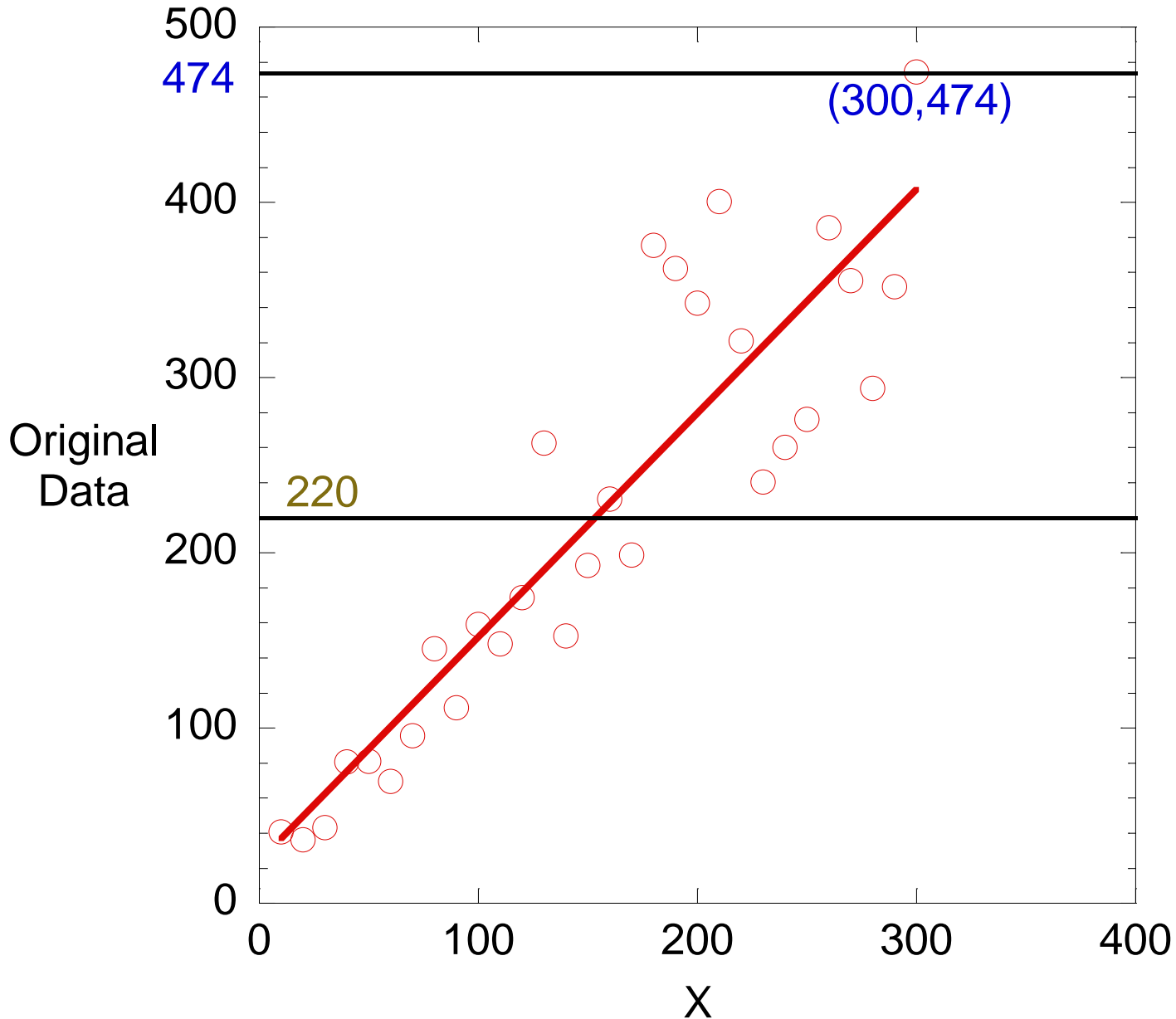
Regression of Y on X



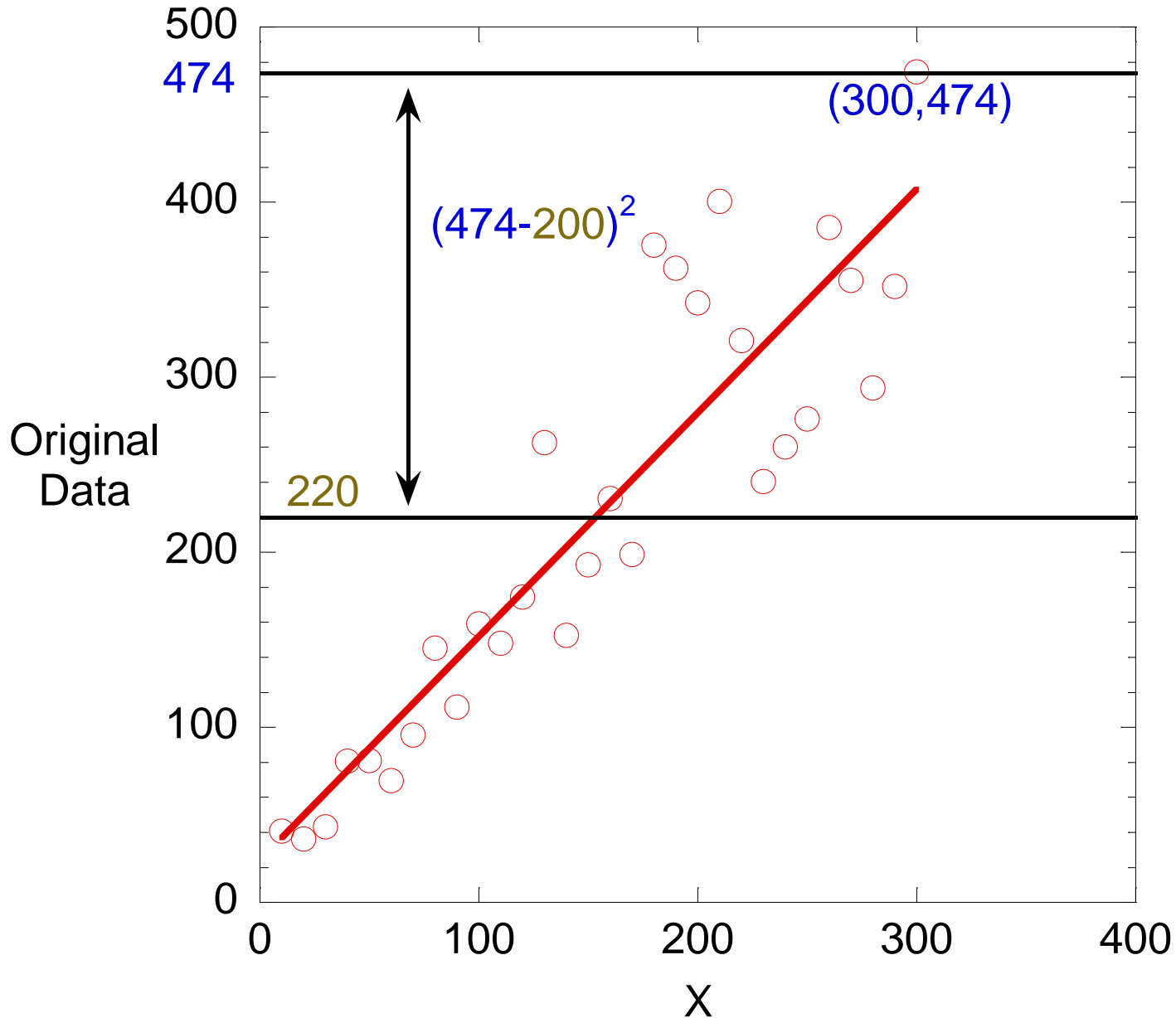
Mean Value of Y



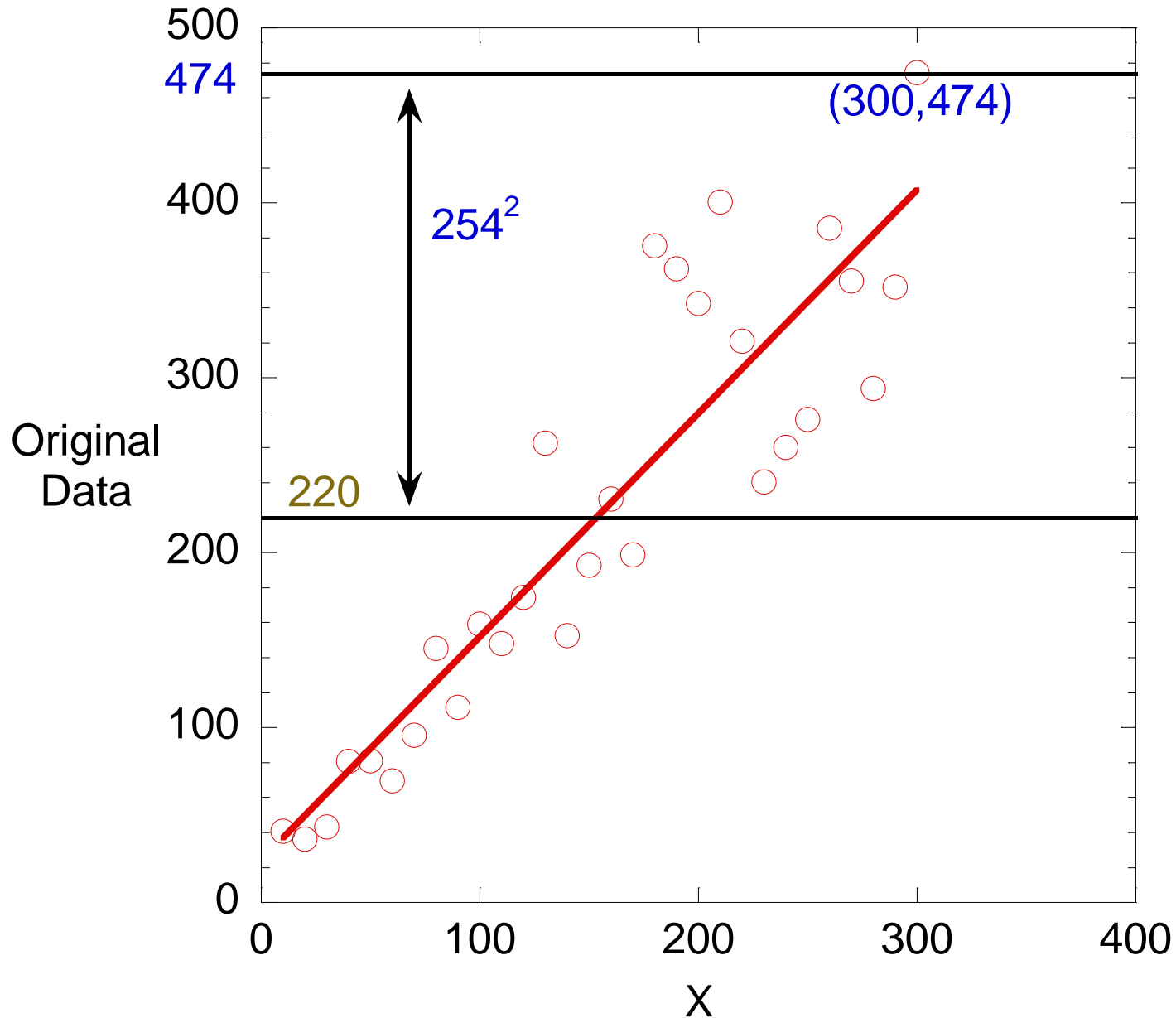
When X is 300 We Observe Y to be 474



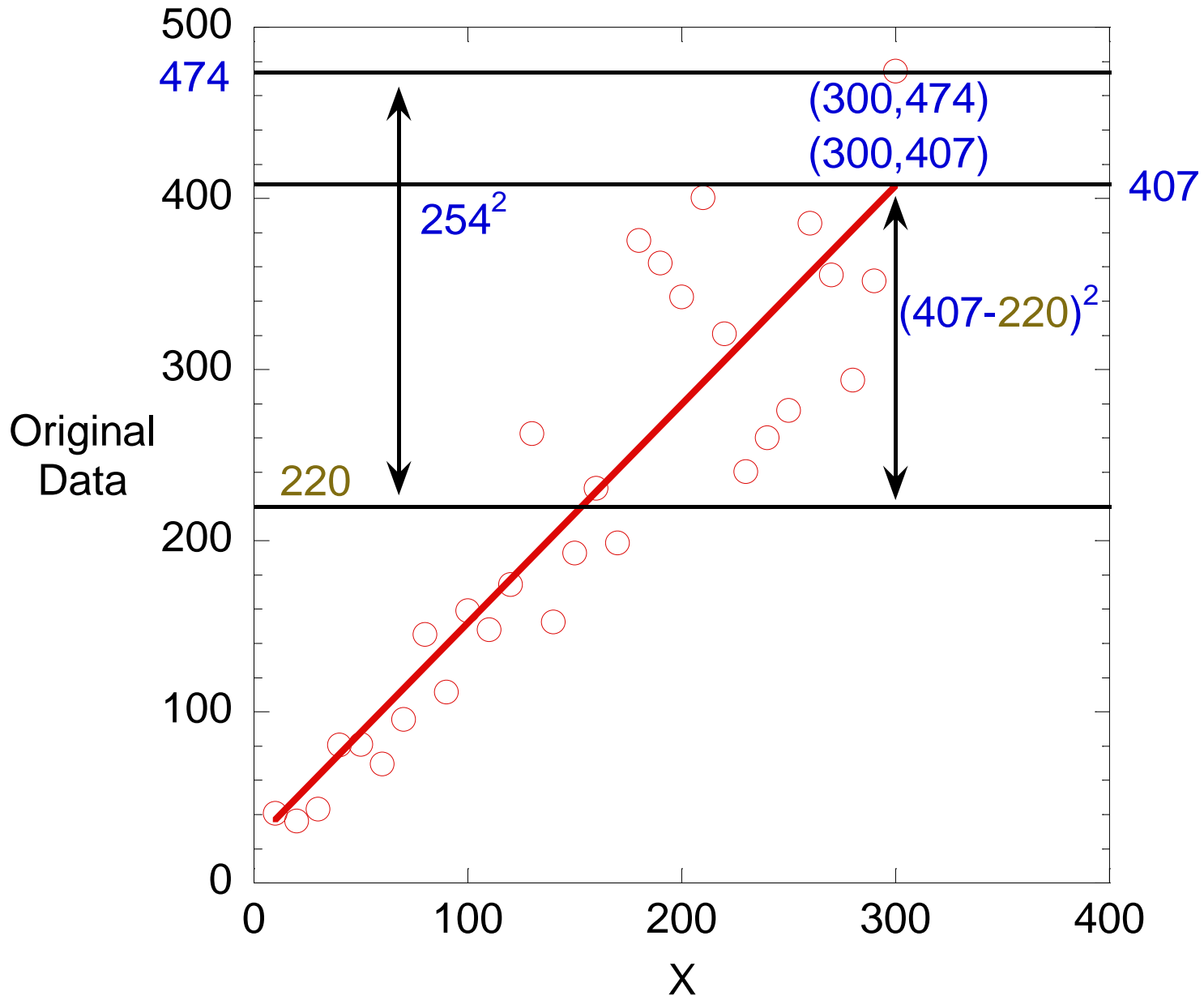
Observed Values Give us Total Variability



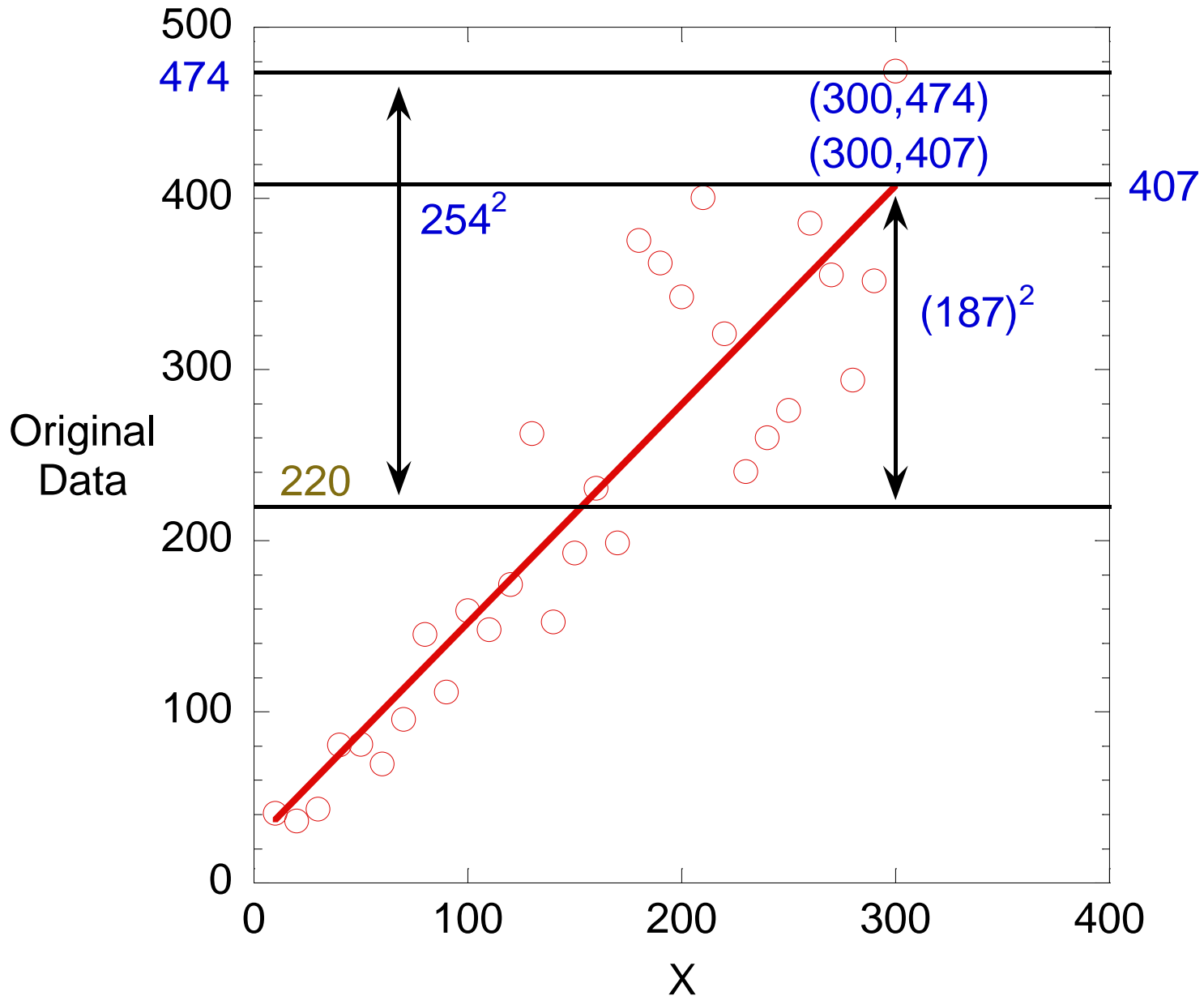
Observed Values Give us Total Variability



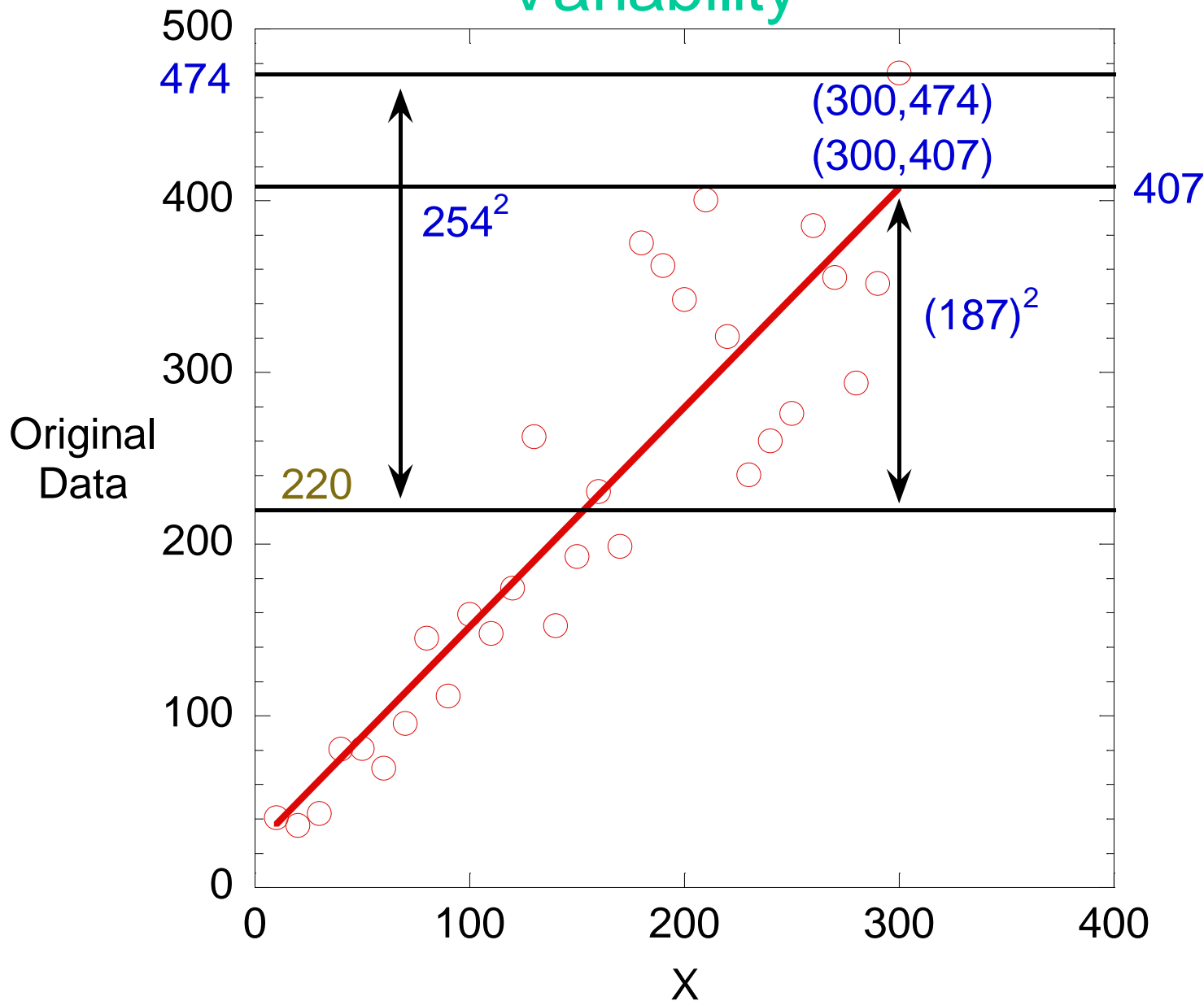
When X is 300 We Predict Y to be 407



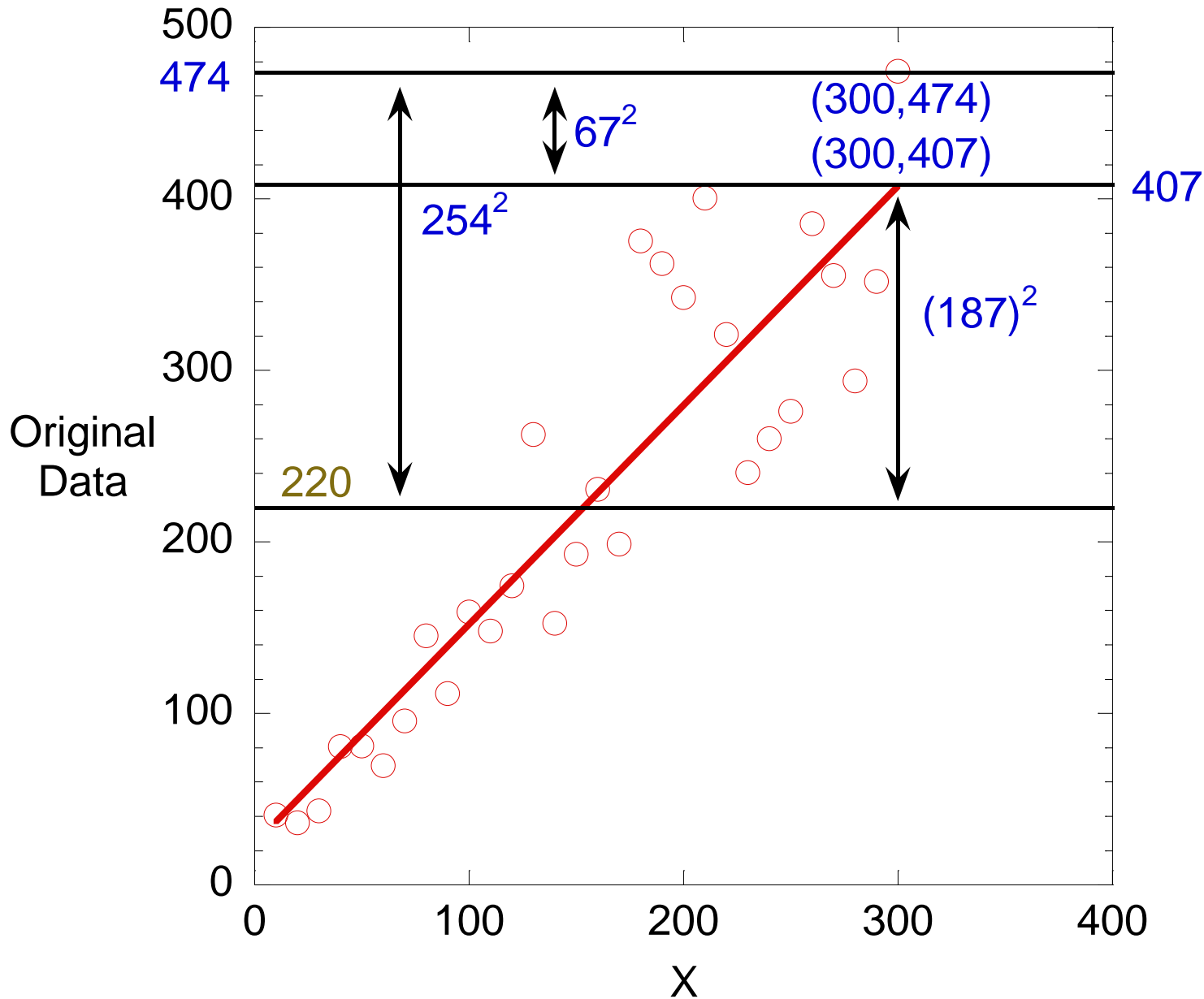
When X is 300 We Predict Y to be 407



Predicted Values Explain Some of the Variability



Some Variability Remains Unexplained



Regression

- Total Variability

$$\sum_{i=1}^N (y_i - \bar{y})^2 \quad \text{Observed-mean}$$

- Explained (Regression) Variability

$$\sum_{i=1}^N (\hat{y}_i - \bar{y})^2 \quad \text{Predicted-mean}$$

- Unexplained (Residual) Variability

$$\sum_{i=1}^N (\hat{y}_i - y_i)^2 \quad \text{Predicted-observed}$$

ANOVA Table from Regression

Source	Df	Sum of Sq	Mean Sq	F Value	Pr(F)
Regression	1	366,645	366,645	122	<0.0001
Residuals	28	83,902	2,997		
Total	29	450,547			

Regression

- Total Variability $\sum_{i=1}^N (y_i - \bar{y})^2 = 450,547$
- Regression Variability $\sum_{i=1}^N (\hat{y}_i - \bar{y})^2 = 366,645$
- Residual Variability $\sum_{i=1}^N (\hat{y}_i - y_i)^2 = 83,902$

Source	Df	Sum of Sq	Mean Sq	F Value	Pr(F)
Regression	1	366,645	366,645	122	<0.0001
Residuals	28	83,902	2,997		
Total	29	450,547			

ANOVA Table from Regression

Source	Df	Sum of Sq	Mean Sq	F Value	Pr(F)
Regression	1	366,645	366,645	122	<0.0001
Residuals	28	83,902	2,997		
Total	29	450,547			

Mean Square Regression

Mean Square Residual

$$\frac{366,645}{2,997} = 122$$

ANOVA Table from Regression

Source	Df	Sum of Sq	Mean Sq	F Value	Pr(F)
Regression	1	366,645	366,645	122	<0.0001
Residuals	28	83,902	2,997		
Total	29	450,547			

Mean Square Regression

Mean Square Residual

$$\frac{366,645}{2,997} = 122$$

$$F(122,1,28) < 0.0001$$

Variability in Regression

- Distance between **mean** and:
 - **Observed** values
 - Total Variability
 - **Predicted** values
 - Explained (Regression) Variability
- Distance between **observed** and **predicted** values
 - Unexplained (Residual) Variability

Reduction in Variance

r^2 , Coefficient of Determination

$$r^2 = \frac{\text{Explained Variance}}{\text{Total Variance}}$$

$$r^2 = \frac{\sum (\text{Predicted} - \text{Mean})^2 / n - 1}{\sum (\text{Observed} - \text{Mean})^2 / n - 1}$$

$$r^2 = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2 / n - 1}{\sum_{i=1}^n (y_i - \bar{y})^2 / n - 1} = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$

Reduction in Variance

r^2 , Coefficient of Determination

$$r^2 = \frac{\text{Explained Variance}}{\text{Total Variance}}$$

$$r^2 = \frac{\sum (\text{Predicted} - \text{Mean})^2 / n - 1}{\sum (\text{Observed} - \text{Mean})^2 / n - 1}$$

$$r^2 = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2 / n - 1}{\sum_{i=1}^n (y_i - \bar{y})^2 / n - 1} = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$

$$r^2 = \frac{366,645}{450,547} = 0.81$$

Non-Parametric Tests

Uses of Non-Parametric Tests

- Used for
 - Nominal
 - Ordinal data
- Can be used on data measured on
 - Interval
 - Ratio scales
 - When used, the tests used are generally *not* the tests listed below.

Non-Parametric Tests

Nominal Data

- Independent Samples
 - Pearson's Chi squared
 - Binomial test
 - Fisher's Exact test
- Related Samples
 - Fisher's Exact test
 - McNemar's chi-square test

Non-Parametric Tests

Ordinal Data

- Independent Samples
 - Mann-Whitney U test
 - Median test
 - Kruskal-Wallis one-way ANOVA by ranks
 - Kolmogorov-Smirnov Test
- Matched Samples
 - Sign Test
 - Wilcoxon test

Non-Parametric Tests for Association

Categorical Independent, Nominal or Ordinal
Dependent

<u>Dependent</u>	<u>Independent</u>	<u>Samples</u>	<u>Test</u>
Nominal	Categorical	Independent	Chi-square
			Binomial test
			Fisher's Exact test
Ordinal	Categorical	Related	McNemar's chi square
		Independent	Mann-Whitney U
			Median test
			Kruskal-Wallis
			Kolmogorov-Smirnov
Matched	Sign test		
			Wilcoxon test

Pearson's Chi-Square

- Outcome (dependent)
 - Nominal
- Predictor (independent)
 - Nominal
- Groups
 - Independent
 - Two or more

Pearson's Chi-Square

- Limitations
 - Outcome nominal, i.e. not ordered
 - Can be used with ordinal (order) outcomes but with loss of power
 - Expected cell frequencies can not be very small
 - Generally must be greater than five

Binomial Test

- Outcome (dependent)
 - Nominal
- Predictor (independent)
 - Nominal
- Groups
 - One Group

Binomial Test

- Limitations
 - Outcome nominal, i.e. not ordered
 - Can be used with ordinal (order) outcomes but with loss of power
 - Only one group
 - N.b. can be extended but . . .

Fisher's Exact Test

- Outcome (dependent)
 - Nominal
- Predictor (independent)
 - Nominal
- Groups
 - Independent
 - Two groups
 - Has been extended to more than two groups
- Can deal with small (even zero) expected cell frequencies

Fisher's Exact Test

- Limitations
 - Outcome nominal, i.e. not ordered
 - Can be used with ordinal (order) outcomes but with loss of power
 - Generally requires computer – especially for more than two groups
 - Some say it is too conservative

Fisher's Exact Test

- Limitations
 - Outcome nominal, i.e. not ordered
 - Can be used with ordinal (order) outcomes but with loss of power
 - Can deal with small (even zero) expected cell frequencies
 - Generally requires computer – especially for more than two groups
 - Some say it is too conservative (religious argument)

McNemar Test

- Outcome (dependent)
 - Nominal
- Predictor (independent)
 - Nominal
- Groups
 - Two Groups
 - Related
 - Matched
 - Repeated measure from same person or sample

McNemar Test

- Limitations
 - Outcome nominal, i.e. not ordered
 - Can be used with ordinal (ordered) outcomes but with loss of power
 - Only discordant pairs contribute to analysis

Mann-Whitney U

- Outcome (dependent)
 - Ordinal
- Predictor (independent)
 - Nominal
- Groups
 - Independent
 - Two or more

Mann-Whitney U

- Limitations
 - A rank based test
 - Need plan to deal with tied ranks
 - Can have problems with small sample sizes

Median Test

- Outcome (dependent)
 - Ordinal
- Predictor (independent)
 - Nominal
- Groups
 - Independent
 - Two or more

Median Test

- Limitations
 - Analysis based on
 - chi-square or
 - Beware of low expected cell frequencies
 - Fisher's exact test
 - Generally requires a computer

Kruskal-Wallis

One-Way Anova by Ranks

- Outcome (dependent)
 - Ordinal
- Predictor (independent)
 - Nominal
- Groups
 - Independent
 - Two or more

Kruskal-Wallis

One-Way Anova by Ranks

- Can have problems with small sample sizes
 - If a sample size in a group is less than 5
 - Use table of exact probabilities

Kolmogorov-Smirnov Test

- Outcome (dependent)
 - Ordinal
- Predictor (independent)
 - Nominal
- Groups
 - Independent
 - Two or more

Kolmogorov-Smirnov Test

- Limitations
 - Known to be conservative if data tested is discrete rather than continuous

Pearson's Chi-square

Karl Pearson 1857-1936



- Developed mathematical methods for studying heredity and evolution
- 1893-1912 18 papers “Mathematical Contributions to the Theory of Evolution”
- Chi-square appeared in 1900 paper

Observed Data

		Prefers Diet Soda	
		Yes	No
Sex	Male	a	b
	Female	c	d

Observed Data

		Prefers Diet Soda	
		Yes	No
Sex	Male	a	b
	Female	c	d
		$a+b+c+d$	

Observed Data

		Prefers Diet Soda		
		Yes	No	
Sex	Male	a	b	a+b
	Female	c	d	
				a+b+c+d

Observed Data

		Prefers Diet Soda		
		Yes	No	
Sex	Male	a	b	a+b
	Female	c	d	c+d
				a+b+c+d

Observed Data

		Prefers Diet Soda		
		Yes	No	
Sex	Male	a	b	a+b
	Female	c	d	c+d
		a+c		a+b+c+d

Observed Data

		Prefers Diet Soda		
		Yes	No	
Sex	Male	a	b	a+b
	Female	c	d	c+d
		a+c	b+d	a+b+c+d

Probability of being Male

		Prefers Diet Soda		
		Yes	No	
Sex	Male	a	b	$(a+b)/(a+b+c+d)$
	Female	c	d	
				$a+b+c+d$

Probability of Preferring Diet Soda

		Prefers Diet Soda	
		Yes	No
Sex	Male	a	b
	Female	c	d
		$(a+c)/(a+b+c+d)$	$a+b+c+d$

Probability of Independent Events

- If two events occur independently
 - X with probability $\frac{1}{2}$
 - Y with probability $\frac{1}{4}$
- The probability of both X and Y occurring is
 - Probability of X * probability of Y

$$p(x | y) = p(x)p(y)$$

$$p(x | y) = (1/2)(1/4) = 1/8$$

Probability of being Male and Preferring Diet Soda

		Prefers Diet Soda		
		Yes	No	
Sex	Male	a	b	$(a+b)/(a+b+c+d)$
	Female	c	d	
		$(a+c)/(a+b+c+d)$		$a+b+c+d$

- Assuming sex and soda preference are independent

$$p(\text{Male} \mid \text{DietSoda}) = \frac{a+b}{a+b+c+d} \bullet \frac{a+c}{a+b+c+d}$$

$$p(\text{Male} \mid \text{DietSoda}) = \frac{(a+b)(a+c)}{(a+b+c+d)^2}$$

Expected Number of Males Preferring Diet Soda

		Prefers Diet Soda		
		Yes	No	
Sex	Male	a	b	$(a+b)/(a+b+c+d)$
	Female	c	d	
		$(a+c)/(a+b+c+d)$		$a+b+c+d$

$$Expected(Male | DietSoda) = (a + b + c + d) \frac{(a + b)(a + c)}{(a + b + c + d)^2}$$

$$Expected(Male | DietSoda) = \frac{(a + b)(a + c)}{a + b + c + d}$$

Expected Number of Females Preferring Diet Soda

		Prefers Diet Soda		
		Yes	No	
Sex	Male	a	b	
	Female	c	d	$(c+d)/(a+b+c+d)$
		$(a+c)/(a+b+c+d)$		$a+b+c+d$

$$Expected(Female | DietSoda) = (a + b + c + d) \frac{(c + d)(a + c)}{(a + b + c + d)^2}$$

$$Expected(Female | DietSoda) = \frac{(c + d)(a + c)}{a + b + c + d}$$

Expected Number of Males Not Preferring Diet Soda

		Prefers Diet Soda		
		Yes	No	
Sex	Male	a	b	$(a+b)/(a+b+c+d)$
	Female	c	d	
		$(b+d)/(a+b+c+d)$		$a+b+c+d$

$$Expected (Male | NotDietSoda) = (a + b + c + d) \frac{(a + b)(b + d)}{(a + b + c + d)^2}$$

$$Expected (Male | NotDietSoda) = \frac{(a + b)(b + d)}{a + b + c + d}$$

Expected Number of Females Not Preferring Diet Soda

		Prefers Diet Soda		
		Yes	No	
Sex	Male	a	b	
	Female	c	d	$(c+d)/(a+b+c+d)$
		$(b+d)/(a+b+c+d)$		$a+b+c+d$

$$Expected(Female | NotDietSoda) = (a + b + c + d) \frac{(c + d)(b + d)}{(a + b + c + d)^2}$$

$$Expected(Female | NotDietSoda) = \frac{(c + d)(b + d)}{a + b + c + d}$$

Chi Squared

- Sum of squared differences between
 - Observed
 - Expected counts
- Divided by Expected count

$$\chi^2 = \sum_{i=1}^n \frac{(O_i - E_i)^2}{E_i}$$

Chi Squared

$$\chi^2 = \sum_{i=1}^n \frac{(O_i - E_i)^2}{E_i}$$

- Contribution for men who prefer diet soda

$$\chi_{Male|DietSoda}^2 = \frac{\left(a - \frac{(a+b)(a+c)}{a+b+c+d} \right)^2}{\frac{(a+b)(a+c)}{a+b+c+d}}$$

What Does Chi Squared Measure?

$$\chi^2 = \sum_{i=1}^n \frac{(O_i - E_i)^2}{E_i}$$

- A sum of variances

What Question Does Chi Squared Ask?

$$\chi^2 = \sum_{i=1}^n \frac{(O_i - E_i)^2}{E_i}$$

- Is the summed variance sufficiently large to reject the null hypothesis?
 - H0: The rows and columns are independent.
- Are the observed cell values sufficiently different from the predicted values (assuming independence) to reject the assumption of independence?

What Question Does Chi Squared Ask?

$$\chi^2 = \sum_{i=1}^n \frac{(O_i - E_i)^2}{E_i}$$

$$\frac{\left(a - \frac{(a+b)(a+c)}{a+b+c+d}\right)^2 + \left(c - \frac{(c+d)(a+c)}{a+b+c+d}\right)^2 + \left(b - \frac{(a+b)(b+d)}{a+b+c+d}\right)^2 + \left(d - \frac{(c+d)(b+d)}{a+b+c+d}\right)^2}{\frac{(a+b)(a+c)}{a+b+c+d}}$$

		Prefers Diet Soda	
		Yes	No
Sex	Male	a	b
	Female	c	d
			a+b+c+d

What Question Does Chi Squared Ask?

How Do We Interpret Result?

		Prefers Diet Soda	
		Yes	No
Sex	Male	a	b
	Female	c	d
			a+b+c+d

$$\frac{\left(a - \frac{(a+b)(a+c)}{a+b+c+d}\right)^2 + \left(c - \frac{(c+d)(a+c)}{a+b+c+d}\right)^2 + \left(b - \frac{(a+b)(b+d)}{a+b+c+d}\right)^2 + \left(d - \frac{(c+d)(b+d)}{a+b+c+d}\right)^2}{\frac{(a+b)(a+c)}{a+b+c+d}}$$

$$\chi^2_{(1,0.05)} = 3.84$$

Pre-Conference stats 07/08/2005:

Evaluation of the presentation (Circle a value that indicates your evaluation of the presentation)

Best ever

Worst ever

10-----9-----8-----7-----6-----5-----4-----3-----2-----1

What I like most about the presentation

What I like least about the presentation

Suggestions:

Questions (If you give me your name, I will try to find you and answer your question):

Basic Truism Redux

Common Statistical Problems

- Study design
- Writing a grant
- Data analysis
- Writing paper

Solution

- Consult a statistician early and often!

Sources of Statistical Help

- John Sorkin, M.D. Ph.D.
 - Chief Biostatistics and Informatics
 - Baltimore VA Medical Center
 - 5-7119
 - JSorkin@grecc.umaryland.edu
- Patricia Langenberg, Ph.D.
 - GCRC Biostatistical Core Director
 - 6-3251
 - PLangenb@umaryland.edu

Sources of Statistical Help

- Min Zhan, Ph.D.
 - Biostatistician, VA Research Service
 - 5-5084, 6-3518
 - MZhan@epi.umaryland.edu
- Samuel Dongmo Ph.D.
 - GCRC Bioinformatics Core Director
 - 8-8007
 - SDong@medicine.umaryland.edu

References

- PDQ Statistics, Norman G, Streiner D, Dekker Inc, Philadelphia, 1986
 - Short, quick, sweet, easy to understand
- Using and Understanding Medical Statistics, Matthews DE, Farewell VT, Krager, New York, 1996
 - Slightly longer, easy to understand, more “formal”
- Biostatistics: A Foundation for Analysis in the Health Sciences, Daniel, WW, John Wiley & Sons 1991
 - Great textbook, comprehensive, easy to understand (for a textbook)

Frederick Mosteller



- **John W. Tukey**
-



Muhammad al-Khwarizmi

Born Approx 783



- Algorithms are named after him
- Scholar, House of Wisdom in Baghdad approx 813.
- Studied algebra, geometry and astronomy.
- Algebra treatise *Hisab al-jabr w'al-muqabala* most famous work.
- First book to be written on algebra.
- Title *al-jabr* (completion) gives us the word "algebra"

1983 USSR postage stamp commemorating the 1200th anniversary of Muhammad al-Khwarizmi³¹³

<http://www-groups.dcs.st-and.ac.uk/~history/Mathematicians/Al-Khwarizmi.html>

- **John W. Tukey**



- **Frederick Mosteller**



Bernoulli Process

Jacob Bernoulli 1654-1705



- Swiss mathematician and physicist
 - Algebra, calculus, series, probability
- Studied philosophy and theology at behest of his parents
 - which he greatly resented
- Studied mathematics and astronomy against the wishes of his parents.
- Developed
 - Bernoulli Process
 - Binomial Distribution

Bernoulli Process

Jacob Bernoulli 1654-1705



- Swiss mathematician and physicist
 - Algebra, calculus, series, probability
- Studied philosophy and theology at behest of his parents
 - which he greatly resented
- Studied mathematics and astronomy against the wishes of his parents.
- Developed
 - Bernoulli Process
 - Binomial Distribution

Bernoulli Process

Jacob Bernoulli 1654-1705




- Swiss mathematician and physicist
 - Algebra, calculus, series, probability
- Studied philosophy and theology at behest of his parents
 - which he greatly resented
- Studied mathematics and astronomy against the wishes of his parents.
- Developed the
 - Bernoulli Process
 - Binomial Distribution

Bernoulli Process

- Each test results in one of two possible outcomes.
- The tests are independent.
- The probability of success, p , is constant.
 - The probability of failure is $1-p$.

Binomial Distribution

 This image cannot currently be displayed.

- Each test results in one of two possible outcomes.
- The tests are independent.
- The probability of success, p , is constant.
 - The probability of failure is $1-p$.

$$\text{prob}(x|n) = {}_n C_x p^x (1-p)^{n-x} \text{ for } x = 0, 1, 2, \dots, N$$


Binomial Distribution

John, what do these mean?

$$\mu = np$$

$$\sigma^2 = np(1 - p)$$

Binomial Distribution

 This image cannot currently be displayed.

- Each test results in one of two possible outcomes.
- The tests are independent.
- The probability of success, p , is constant.
 - The probability of failure is $1-p$.

$$\text{prob}(x|n) = \binom{n}{x} p^x (1-p)^{n-x} \text{ for } x = 0, 1, 2, \dots, N$$

Poisson Process

Siméon Denis Poisson 1781-1840



- Grew up during French Revolution
- Studied to be a surgeon
- Mathematics at École Polytechnique
 - Laplace
 - Lagrange

Poisson Distribution

Limit of Binomial Distribution

$$\text{prob}(x|n) = {}_n C_x p^x (1-p)^{n-x}$$

$$\text{prob}(x|n) = \binom{n}{x} p^x (1-p)^{n-x}$$

$$\text{prob}(x|n) = \left(\frac{n!}{x!(n-x)!} \right) p^x (1-p)^{n-x}$$

Poisson Distribution

Limit of Binomial Distribution

$$\text{prob}(x|n) = \left(\frac{n!}{x!(n-x)!} \right) p^x (1-p)^{n-x}$$

$$\left(\frac{n!}{x!(n-x)!} \right) = \left(\frac{n(n-1)(n-2)\dots(n-x+1)}{x!} \right) \approx \left(\frac{n^x}{x!} \right)$$

Poisson Distribution

Limit of Binomial Distribution

$$\text{prob}(x|n) = \binom{n!}{x!(n-x)!} p^x (1-p)^{n-x}$$

$$\binom{n!}{x!(n-x)!} = \frac{n(n-1)(n-2)\dots(n-x+1)}{x!} \approx \left(\frac{n^x}{x!} \right)$$

Poisson Distribution

Limit of Binomial Distribution

- Let
 - λ = Number of events/unit time interval.
 - T = Total observation time.
 - n = Number of intervals of observation.
- Probability of an event in an interval of observation = $\lambda T/n$
- Probability of no event in an interval of observation =

Poisson Distribution

Limit of Binomial Distribution

$$\text{prob}(x|n) = \binom{n^x}{x!} \left(\frac{\lambda T}{n}\right)^x \left(1 - \frac{\lambda T}{n}\right)^{n-x}$$

$$\left(1 - \frac{\lambda T}{n}\right)^{n-x} \approx \left(1 - \frac{\lambda T}{n}\right)^n = e^{-\lambda T}$$

$$\text{prob}(x|n) = \binom{n^x}{x!} \left(\frac{\lambda T}{n}\right)^x e^{-\lambda T}$$

Poisson Distribution

Limit of Binomial Distribution

 This image cannot currently be displayed.

- Sampling is from an infinite population
 - Or finite population with replacement
 - Practically can be used in finite population without replacement if
 - $n < \text{Population size}/10$

Frederick Mosteller

1916-













Pierre de Fermat



