MATH 425/525

Hao Wang

MATH 425/525

Outline of the Course

1. Data Aanlysis and Computer

data histogram, data mean, data variance

2. Probability Foundation

Events, sample space, counting rules, conditional probability, independence, Bayes' rule, discrete random variables, binomial distribution, continuous random variables, normal distribution, central limit theorem

3. Basic Statistics

sampling distributions, type of estimators,point estimation for large sample case, interval estimation for large sample case

Objective of the Course

This course is the foundation part of the sequence of MATH 425/525, MATH 426/526, and MATH 427/527. This sequence can provide your research with useful

probability and statistics techniques; for example, computer involved data analysis, decision making, making inference, and so on.

Data Analysis

What is statistics ?

Statistics is an area of science concerned with the extraction of information from numerical data and its use in making inference on the population from which the numerical data are obtained.

What is population?

Definition 1.1 A population or sample space is the set representing all measurements of interest to the investigator

What does it mean making inference?

Example 1.1 Suppose that in a company a quality control person wants to make a decision on whether this shipment of 100,000 batteries is qualified to deliver. The lifetime of a battery is the single quality measurement. Therefore,

sample space = { all the lifetimes of this shipment of batteries }

If the lifetime of a battery is longer than 1000 hours, then this battery is qualified. Also if 98% of the batteries in a shipment is qualified, then this shipment of batteries is qualified to deliver. Since we can't check

the batteries one by one, we only can check very limited number of batteries, then we use statistic technique to make inference and decision.

Definition 1.2 a sample is a subset of measurements selected from the population of interest.

Example 1.2 Suppose that the U.S. government wants to change a family income related policy which is based on the average of family incomes in U.S.. Therefore,

sample space = { all the family incomes in U.S. }

Eugene's family incomes is a sample.

Definition 1.3 A population range is an interval from the smallest to the largest measurements in the population. If the population range is decomposed into a group of mutually exclusive categories or subintervals, then the frequency is the number of measurements in each category and the relative frequency is the proportion of measurements in each category.
Definition 1.4 If we equally decompose a population range to get a group of categories or classes, a relative frequency histogram for a data set or population is a bar graph in which the height of the bar represents the relative frequency of a category

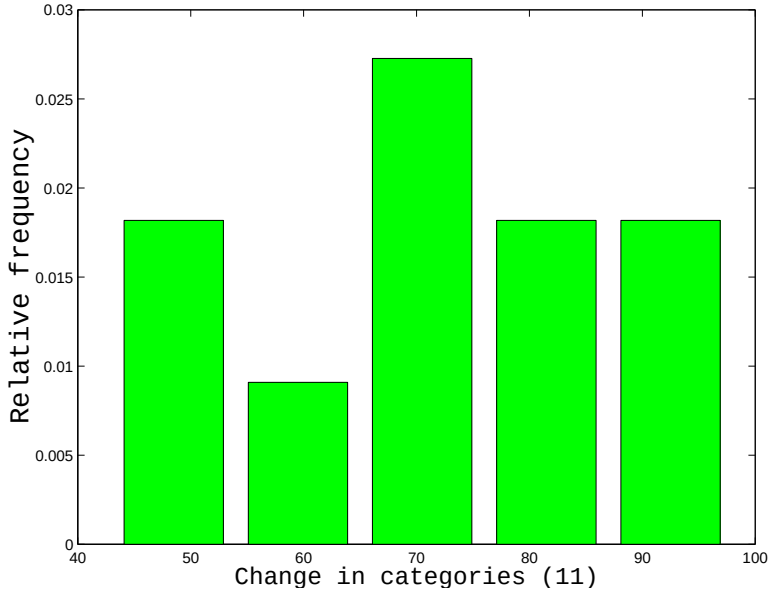
Example 1.3 Suppose that in a class there are 10 students whose final exam scores out of 100 are as follows:

98, 90, 85, 80, 70, 69, 75, 60, 43, 50

If we equally divide the population range $98 - 43 = 55$ into five mutually exclusive categories

$$[43, 54), [54, 65), [65, 76), [76, 87), [87, 98]$$

Then, they are mutually exclusive. The relative frequency histogram is as follows:



Extracting Information from Data

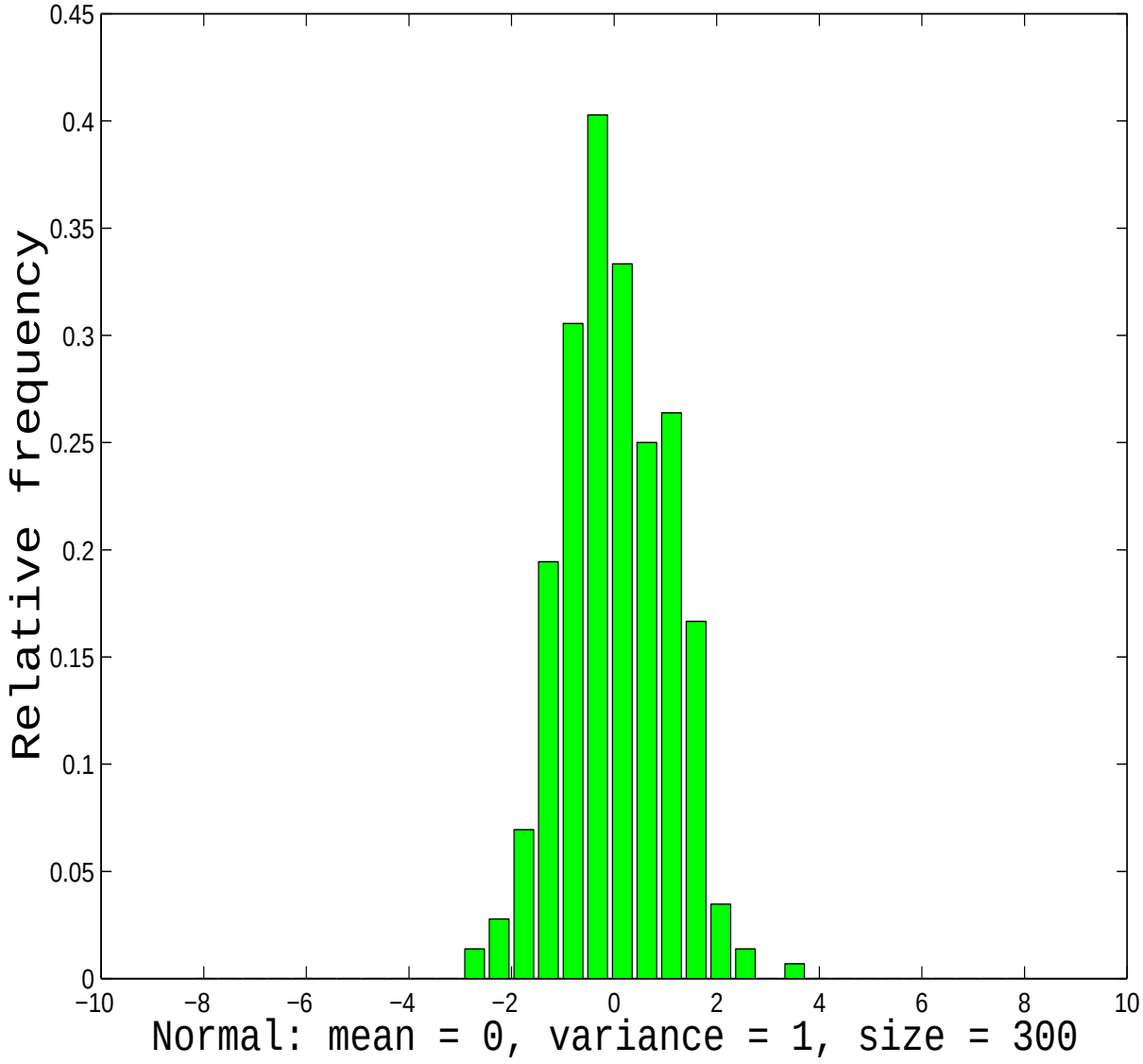
The average of a sample is called sample mean which gives the measure of center of the sample.

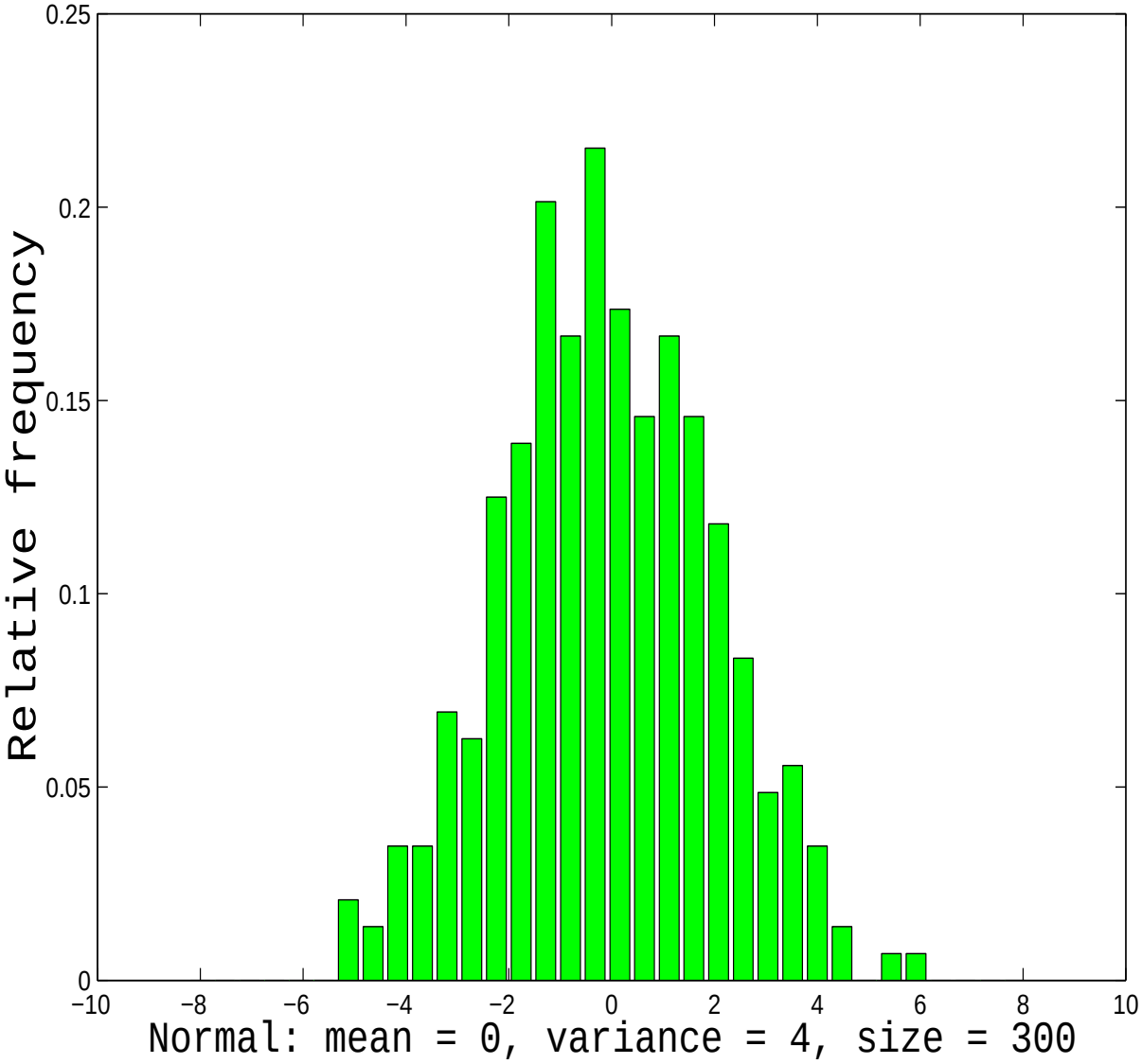
Definition 1.5 Suppose that we have a sample $x_1, \cdots, x_n$ of n measurements. The sample mean of $x_1, \cdots, x_n$ is defined by:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

The following two histograms have same mean, but

their dispersions or variabilities are different. This means that their ways of spreading out are different.




0.25
0.2
0.15
0.1
0.05
0
Relative frequency
−10
−8
−6
−4
−2
0
2
4
6
8
10
Normal: mean = 0, variance = 4, size = 300

Definition 1.6 Suppose that we have a sample $x_1, \cdots, x_n$ of n measurements. The sample variance of $x_1, \cdots, x_n$ is defined by:

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} \{x_i - \bar{x}\}^2.$$

$s = \sqrt{s^2}$ is called the sample standard deviation.
s^2 is a measure of sample dispersion around its mean.

In the special case that n is the population size. This means that $\{x_1, \cdots, x_n\}$ is all the measurements of the population, then we have special notation for population mean:

$$\mu = \frac{1}{n} \sum_{i=1}^{n} x_i,$$

population variance:

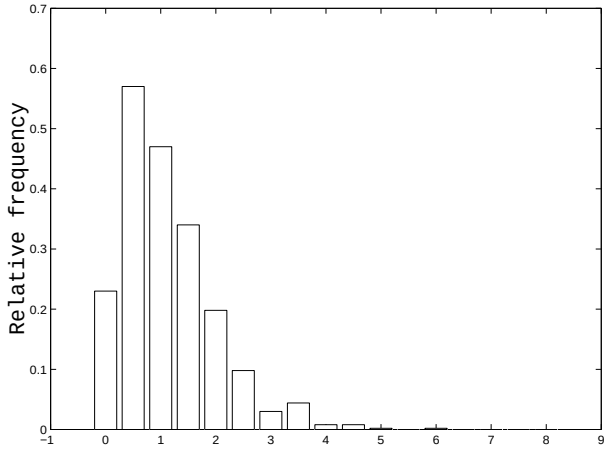
$$\sigma^2 = \frac{1}{n-1} \sum_{i=1}^{n} \{x_i - \mu\}^2.$$

Tchebysheff's Theorem Given a number $k \geq 1$ and a population with n measurements, at least $[1 - (1/k^2)]$ of the measurements will lie within k standard deviations of their mean.

A Simplified Tchebysheff's Theorem
At least 3/4 of th measurements lie in the interval $(\mu - 2\sigma, \mu + 2\sigma)$.
At least 8/9 of th measurements lie in the interval

$(\mu - 3\sigma,\ \mu + 3\sigma)$.

The Tchebysheff's Theorem applies to any shape of relative frequency histogram (r.f.h.), so it is very conservative. The following r.f.h. is not symmetric (skewed).



Example 1.4

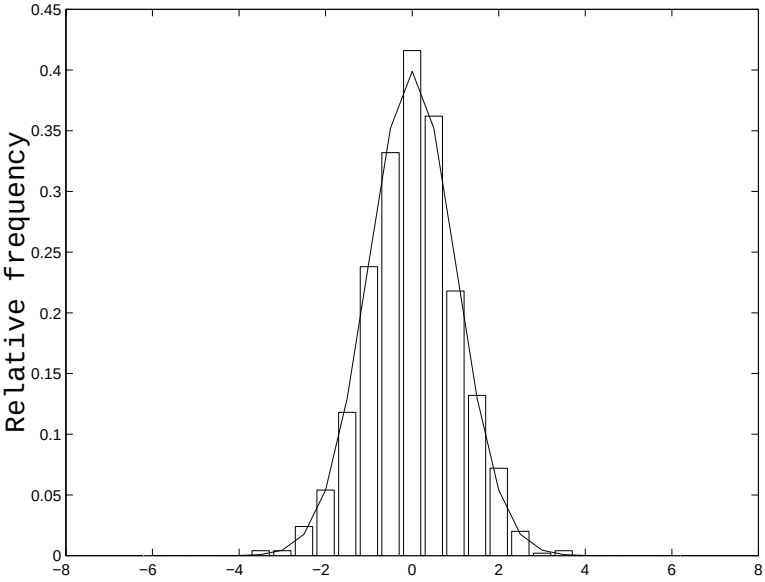
Suppose that the population histogram is symmetric about its mean and the mean and variance of the population of $n = 108$ measurements are 60 and 100, respectively. Use Tchebysheff's theorem to prove that at most 13 measurements are greater than 80.

Proof: In Tchebysheff's theorem, we take $k = 2$, then at least 3/4 of the measurements lie within $[60 - 2 * 10, 60 + 2 * 10]$. In other words, at most 1/4 or 27 of the measurements lie outside of $[60 - 2 * 10, 60 + 2 * 10]$. Since the sample histogram is symmetric about its mean, at most 13 of the measurements are greater than $60+2*10 = 80$.

If the shape of the r.f.h. is symmetric and bell shaped, (This is just the normal distribution), then we have a better estimation rule. Precisely, if the r.f.h broken line is very close to the curve of function:

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

then we say that the r.f.h. is symmetric and bell shaped. The following figure show you this meaning.



For a symmetric and bell shaped r.f.h., we have following better estimation rule.

Empirical Rule

Given a population of measurements that is approximately bell shaped, then we have the following estimations:

The interval $(\mu - \sigma, \mu + \sigma)$ contains approximately 68% of the measurements.

The interval $(\mu-2\sigma, \mu+2\sigma)$ contains approximately 95% of the measurements.
The interval $(\mu-3\sigma, \mu+3\sigma)$ contains all or almost all of the measurements.

Example 1.5

The number of television viewing hours per household and the prime viewing times are two factors that affect television advertising income. A random sample of 25 households in a particular viewing area produced the following estimates of viewing hours per household:

6.5	**8.0**	**4.0**	**5.5**	**6.0**
3.0	**6.0**	**7.5**	**15.0**	**12.0**
5.0	**12.0**	**1.0**	**3.5**	**3.0**
7.5	**5.0**	**10.0**	**8.0**	**3.5**
9.0	**2.0**	**6.5**	**1.0**	**5.0**

Find the percentage of the viewing hours per household that falls into the interval $(\bar{x}-2s, \bar{x}+2s)$. Compare with the corresponding percentage given by the Empirical Rule.

Solution **Here** $n = 25$, $\sum_{i=1}^{25} x_i = 155.5$, $\sum_{i=1}^{25} x_i^2 = 1260.75$. **Then**

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{25}\sum_{i=1}^{25} x_i = \frac{155.5}{25} = 6.22$$

$$s = \sqrt{\frac{\sum_{i=1}^{n}\{x_i - \bar{x}\}^2}{n-1}} = \sqrt{\frac{\sum_{i=1}^{25} x_i^2 - \frac{(\sum_{i=1}^{25} x_i)^2}{n}}{n-1}}$$

$$= \sqrt{\frac{1260.75 - \frac{(155.5)^2}{25}}{24}} = 3.497$$

We can get $(\bar{x} - 2s, \bar{x} + 2s) = (6.22 - 6.994, 6.22 + 6.994) = (-0.774, 13.214)$**. From the original data, 24 measurements or (24/25)100 = 96% of the measurements fall in this interval. This is close to the 95% result of Empirical rule.**