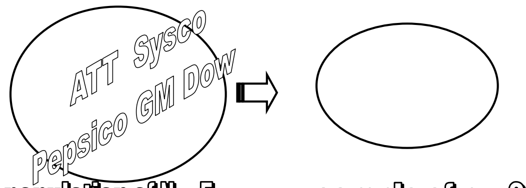
Raoul LePage
Professor
STATISTICS AND PROBABILITY
www.stt.msu.edu/~lepage
click on STT461_Sp05

Slides 38-41 revised 1-7-05.

1

THE GREAT TRICK OF STATISTICS

The overwhelming majority of samples of n from a population of N can stand-in for the population.

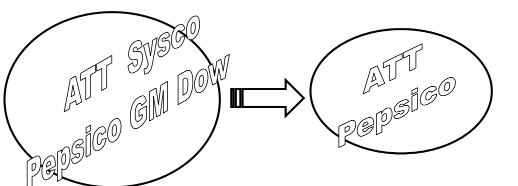


population of N = 5 **sample of n = 2**

2

THE GREAT TRICK OF STATISTICS

The overwhelming majority of samples of n from a population of N can stand-in for the population.



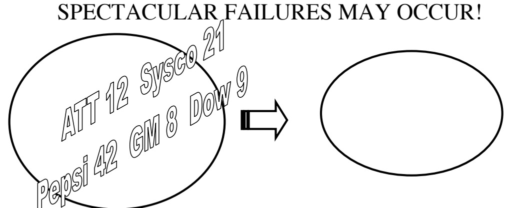
population of N = 5 **sample of n = 2**

3

GREAT TRICK : SOME CAVEATS

For a few characteristics at a time, such as profit, sales, dividend. Sample size n must be "large."

SPECTACULAR FAILURES MAY OCCUR!



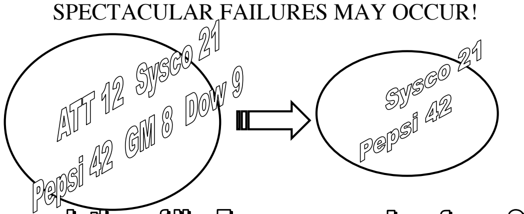
population of N = 5 **sample of n = 2**

4

GREAT TRICK : SOME CAVEATS

For a few characteristics at a time, such as profit, sales, dividend. Sample size n must be "large."

SPECTACULAR FAILURES MAY OCCUR!

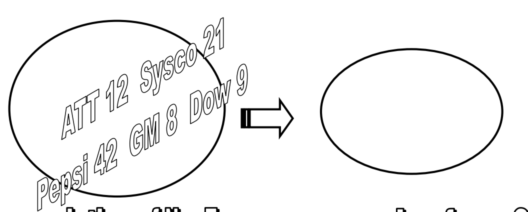


population of N = 5 **sample of n = 2**

5

HOW ARE WE SAMPLING ?

With-replacement
vs without replacement.

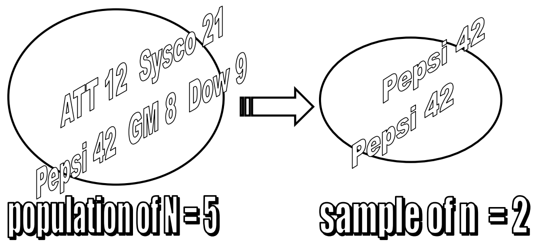


population of N = 5 **sample of n = 2**

6

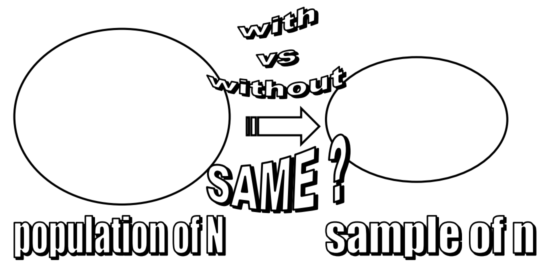
HOW ARE WE SAMPLING ?

With-replacement
vs without replacement.



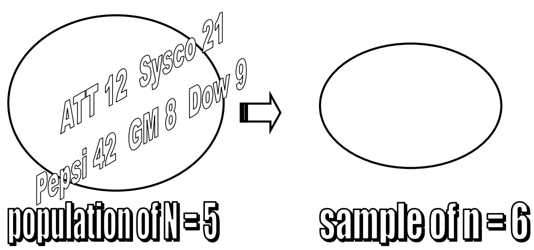
DOES IT MAKE A DIFFERENCE ?

Rule of thumb: With and without replacement are about the same if root [(N-n) /(N-1)] ~ 1.



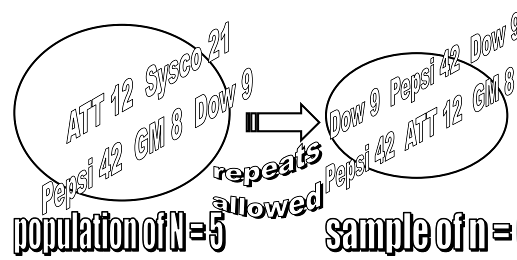
UNLIMITED SAMPLING

WITH-replacement samples have no limit to the sample size n.



UNLIMITED SAMPLING

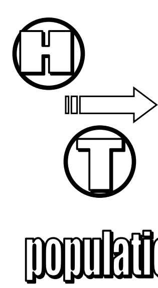
WITH-replacement samples have no limit to the sample size n.



TOSS COIN 100 TIMES

TOSS COIN 100 TIMES

H , T , H , T , T , H , H , H , H , T , T , H , H , H ,
H , H , H , H , T , T , H , H , H , H , H , H , H , H ,
H , T , H , T , H , H , H , H , T , T , H , H , T , T ,
T , T ,T , T , T , H , H , T , T , H , T , T , H , H , H ,
T , H , T , H , T , T , H , H , T ,T , T , H , T , T , T ,
T , T , H , H , T , H , T , T , T , T , H , H , T , T ,
H , T , T , T , T , H , T , H , H , T , T , T , T , T



sample of n = 100
with-replacement
49 H and 51 T

POPULATION CLONING, MANY COINS SAME AS 2

TOSS COIN 100 TIMES

H

H

T

T

T , H , T , T , H , H , H , H , T , T , H , H , H ,
H , H , H , H , T , T , H , H , H , H , H , H , H , H ,
H , T , H , T , H , H , H , H , T , T , H , H , T , T ,
T , T ,T , T , T , H , H , T , T , H , T , T , H , H , H ,
T , H , T , H , T , T , H , H , T ,T , T , H , T , T , T ,
T , T , H , H , T , H , T , T , T , T , H , H , T , T ,
H , T , T , T , T , H , T , H , H , T , T , T , T , T

population of N

sample of n = 100
with-replacement
49 H and 51 T

HOW MANY SAMPLES ARE THERE ?

$\frac{5!}{2!\,3!}$ without replacement 2 from 5

10

5^2 with replacement 2 from 5

25

2 from 5

$\frac{500!}{30!\,470!}$ without replacement 30 from 500

1445259588120573502872111127079547366704414552400

500^{30} with replacement 30 from 500

931322574615478515625000000000000000000000000000000000000\
000000000000000000000000

$\frac{\text{30 from 500 with replacement}}{\text{30 from 500 without replacement}}$

6.44398×10^{32}

HOW MANY SAMPLES ARE THERE ?

$\frac{5!}{2!\,3!}$ without replacement 2 from 5

10

5^2 with replacement 2 from 5

25

$\frac{500!}{30!\,470!}$ without replacement 30 from 500

1445259588120573502872111127079547366704414552400

30 from 500

500^{30} with replacement 30 from 500

931322574615478515625000000000000000000000000000000000000\
000000000000000000000000

$\frac{\text{30 from 500 with replacement}}{\text{30 from 500 without replacement}}$

6.44398×10^{32}

HOW MANY SAMPLES ARE THERE ?

$\frac{5!}{2!\,3!}$ without replacement 2 from 5

10

5^2 with replacement 2 from 5

25

$\frac{500!}{30!\,470!}$ without replacement 30 from 500

1445259588120573502872111127079547366704414552400

500^{30} with replacement 30 from 500

931322574615478515625000000000000000000000000000000000000\
000000000000000000000000

$\frac{\text{30 from 500 with replacement}}{\text{30 from 500 without replacement}}$

6.44398×10^{32}

there are far more samples with-replacement

CORRECTION TO PAGE 17 OF TEXT

They would have you believe
the population is {8, 9, 12, 42}
and the sample is {42}.
A SET is a collection of distinct entities.

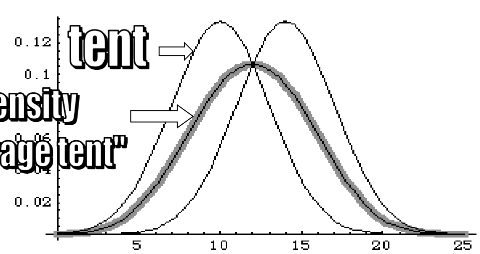
WE SAMPLE COMPANIES
NUMBERS COME WITH THEM

ATT 12 IBM 42 AAA 9
Pepsi 42 GM 8 Dow 9

Pepsi 42
Pepsi 42

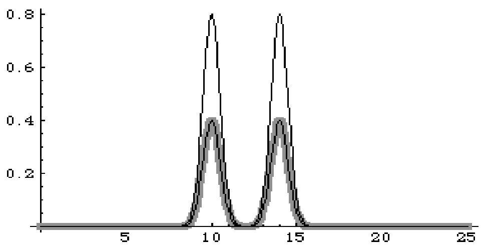
SMOOTHING DATA SO YOU CAN SEE IT

Plot the average heights of tents placed at {10, 14}. Each tent has integral 1, as does their average.



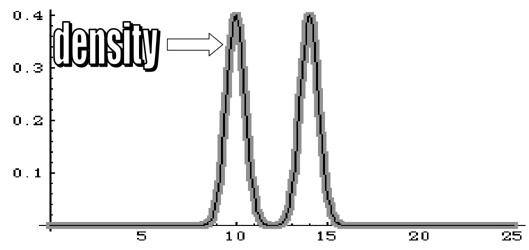
NARROWER TENTS = MORE DETAIL

Making the tents narrower isolates different parts of the data and reveals more detail.



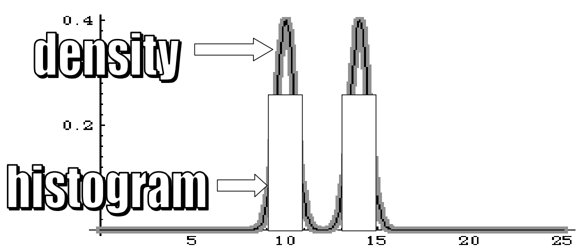
THE DENSITY BY ITSELF

With narrow tents.



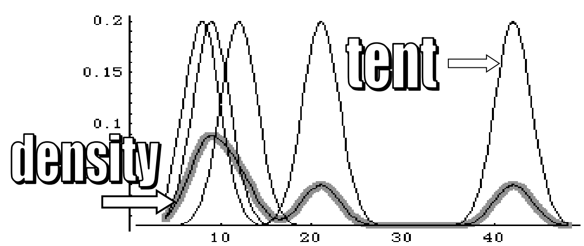
DENSITY OR HISTOGRAM ?

Histograms lump data into categories (the black boxes), not as good for continuous data.



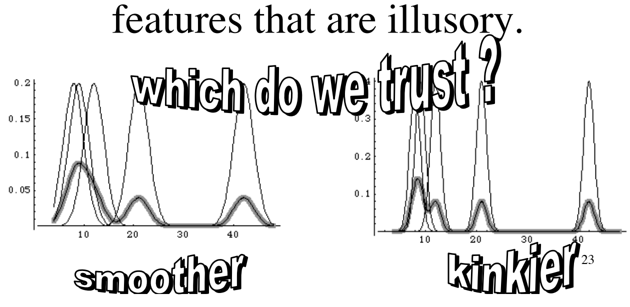
DENSITY FOR { 12, 21, 42, 8, 9 }

Plot of average heights of 5 tents placed at data {12, 21, 42, 8, 9}.



IS DETAIL ILLUSORY ?

Narrower tents operate at higher resolution but they may bring out features that are illusory.

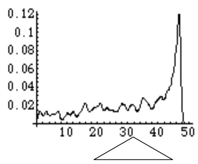


THE MEAN OF A DENSITY IS THE SAME AS THE MEAN OF THE DATA FROM WHICH IT IS MADE.

BEWARE OVER-FINE RESOLUTION

Population of N = 500 compared with two samples of n = 30 each.

POP mean = 32.02



population of N = 500

with 2 samples of n = 30

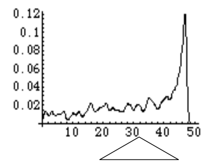
BEWARE OVER-FINE RESOLUTION

Population of N = 500 compared with two samples of n = 30 each.

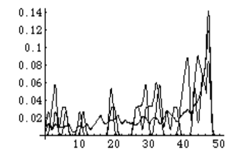
POP mean = 32.02

sample means are close

SAM1 mean = 33.03
SAM2 mean = 30.60



densities not good at fine resolution



population of N = 500

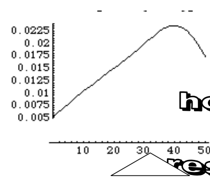
with 2 samples of n = 30

WE DO BETTER AT COARSE RESOLUTION

The same two samples of n = 30 each from the population of 500.

SAM1 mean = 33.03
SAM2 mean = 30.60

POP mean = 32.02



how about coarse resolution ?

population of N = 500

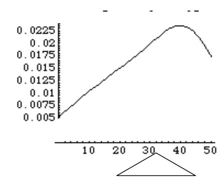
with 2 samples of n = 30

WE DO BETTER AT COARSE RESOLUTION

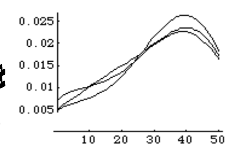
The same two samples of n = 30 each from the population of 500.

SAM1 mean = 33.03
SAM2 mean = 30.60

POP mean = 32.02



good at coarse resolution



population of N = 500

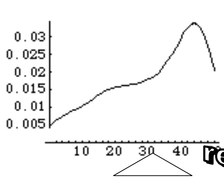
with 2 samples of n = 30

HOW ABOUT MEDIUM RESOLUTION ?

The same two samples of n = 30 each from the population of 500.

SAM1 mean = 33.03
SAM2 mean = 30.60

POP mean = 32.02



medium resolution ?

population of N = 500

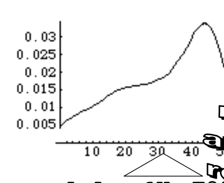
with 2 samples of n = 30

HOW ABOUT MEDIUM RESOLUTION ?

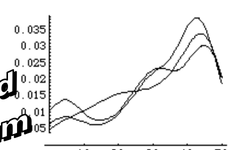
The same two samples of n = 30 each from the population of 500.

SAM1 mean = 33.03
SAM2 mean = 30.60

POP mean = 32.02



not good at medium resolution



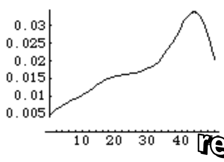
population of N = 500

with 2 samples of n = 30

SAMPLING ONLY 600 FROM 500 MILLION ?

A sample of only n = 600 from a population of N = 500 million. (medium resolution)

POP mean = 32.02



medium resolution ?

large sample of n = 600 ?

population of N = 500,000 with a sample of n = 600

SAMPLING ONLY 600 FROM 500 MILLION ?

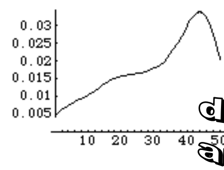
A sample of only n = 600 from a population of N = 500 million. (MEDIUM resolution)

POP mean = 32.02

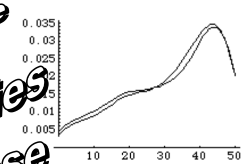
mean very close

sample of n = 600

sample mean = 32.84



densities are close



population of N = 500,000 with a sample of n = 600

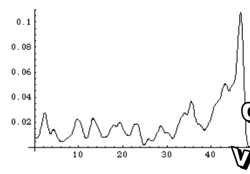
SAMPLING ONLY 600 FROM 500 MILLION ?

A sample of only n = 600 from a population of N = 500 million. (FINE resolution)

POP mean = 32.02

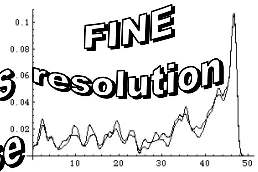
sample of n = 600

sample mean = 32.84



densities very close

FINE resolution



population of N = 500,000 with a sample of n = 600

THE ROLE OF RANDOM SAMPLING

IF THE OVERWHELMING MAJORITY OF SAMPLES ARE "GOOD SAMPLES" THEN WE CAN OBTAIN A "GOOD" SAMPLE BY RANDOM SELECTION.

HOW TO SAMPLE RANDOMLY ?

SELECTING A LETTER AT RANDOM

With-replacement:

a = 00-02 b = 03-05 …. z = 75-77

From Table 14 pg. 869:

1559 9068 9290 8303 etc…

15 59 90 68 etc… (split into pairs)

we have 15 = f , 59 = t, 90 = none, etc…

(for samples without replacement just pass over any duplicates).

TALKING POINTS

1. The Great Trick of Statistics.
   - 1a. The overwhelming majority of all samples of n can "stand-in" for the population to a remarkable degree.
   - 1b. Large n helps.
   - 1c. Do not expect a given sample to accurately reflect the population in many respects, it asks too much of a sample.
2. The Law of Averages is one aspect of The Great Trick.
   - 2a. Samples typically have a mean that is close to the mean of the population.
   - 2b. Random samples are nearly certain to have this property since the overwhelming majority of samples do.
3. A density is controlled by the width of the tents used.
   - 3a. Small samples zero-in on coarse densities fairly well .
   - 3b. Samples in hundreds can perform remarkably well.
   - 3c. Histograms are notoriously unstable but remain popular.
4. Making a density from two to four values; issue of resolution.
5. With-replacement vs without; unlimited samples.
6. Using Table 14 to obtain a random sample.

The Great Trick is far more powerful than we have seen.
A typical sample closely estimates such things as a population mean or the shape of a population density.
But it goes beyond this to reveal how much variation there is among sample means and sample densities.

A typical sample not only estimates population quantities.
It estimates the sample-to-sample variations of its own estimates.

EXAMPLE : ESTIMATING A MEAN

The average account balance is \$421.34 for a random with-replacement sample of 50 accounts.

We estimate from this sample that the average balance is \$421.34 for all accounts.

From this sample we also estimate and display a "margin of error"

\$421.34 +/- \$65.22 = $\bar{X} \pm 1.96 \frac{s}{\sqrt{n}}$

s denotes "sample standard deviation"

SAMPLE STANDARD DEVIATION

$$s = \sqrt{\frac{\sum_{i=1}^{i=n} (X_i - \bar{X})^2}{n-1}}$$

NOTE: Sample standard deviation s may be calculated in several equivalent ways, some sensitive to rounding errors, even for n = 2.

EXAMPLE : MARGIN OF ERROR CALCULATION

The following margin of error calculation for n = 4 is only an illustration. A sample of four would not be regarded as large enough.
Profits per sale = {12.2, 15.3, 16.2, 12.8}.
Mean = 14.125, s = 1.92765, root(4) = 2.
Margin of error = +/- 1.96 (1.92765 / 2)
Report: 14.125 +/- 1.8891.

A precise interpretation of margin of error will be given later in the course, including the role of 1.96. The interval 14.125 +/- 1.8891 is called a "95% confidence interval for the population mean."

We used: $(12.2-14.125)^2 + (15.3-14.125)^2 + (16.2-14.125)^2 + (12.8-14.125)^2 = 11.1475$.

EXAMPLE : ESTIMATING A PERCENTAGE

A random with-replacement sample of 50 stores participated in a test marketing. In 39 of these 50 stores (i.e. 78%) the new package design outsold the old package design.

We estimate from this sample that 78% of all stores will sell more of new vs old.

We also estimate a "margin of error"

+/- 11.6%

percentage is the mean for 100 = new sale, 0 = old sale