Statistical Sampling

Summer 2003

STATISTICAL SAMPLING: An Example

NEXNet is a relatively small but aggressive player in the telecommunications market in the mid-Atlantic region of the US. It is now considering a move into the Boston area.

NEXNet would like to *estimate* the average monthly phone bill in the communities of Weston, Wayland, and Sudbury, by conducting a phone survey. As an enticement for people to participate in the survey, NEXNet will offer discount coupons on certain products to survey participants.

- How many households should NEXNet plan to survey (successfully) in order to effectively estimate the average phone bill in these three communities?

- How should NEXNet analyze the survey results?

Outline

- Random sample
- The sample mean and the sample standard deviation
- The distribution of the sample mean
- Confidence interval estimation.
- Sample size design

Random Sample

- Population: set of all units of interest

- Sample: a subset of the population
- Random Sample: a sample collected in such a way that every member in the population is equally likely to be selected.

Our Goal:
Make inferences, i.e., estimates, predictions, etc. about a population based on information from a sample.

In particular, we want to estimate the population mean μ , and the population standard deviation σ.

Examples of Statistical Sampling

- Marketing: Determine household income of consumers
- Manufacturing: Determine the fraction of defects in a batch
- Polling: Determine the proportion of population that favors a candidate
- Other Examples?

A Failed Survey

Example: 1936 U.S. presidential election, Alf Landon vs. Franklin Roosevelt

- October 1936, Literary Digest conducted the largest poll in history: 10 million voter surveys mailed out. They had correctly predicted the winner since 1916 elections.
- The 2.4 million who completed the survey predicted that Landon would win by 57% to 43%.
- One month later, Roosevelt was re-elected with the largest majority in U.S. history.
- Results: Roosevelt 62%   Landon 38%
- The magazine went bankrupt soon after.

What went wrong?

Biased Sampling

- Names gathered from mailing lists, subscriptions, and telephone books
- Only 1 in 4 households had phones, biased toward the wealthy (who supported Landon whereas the poor supported Roosevelt)
- Only 20% of surveys were returned (non-response bias)
- At the same time, George Gallup polled 3000 Literary Digest readers and correctly predicted the results. He also polled 50,000 potential voters in a less biased sample and predicted Roosevelt would get 56%.
- A larger biased sample does not make a better sample!

A Financial Example

- Imagine that you receive an email from an investment firm offering advice on winning stocks, including a “free sample” stock pick
- The stock goes up that week
- You receive a second email naming a second stock that will go up in the next week
- It goes up
- A third email offers a third stock which goes up
- The fourth email solicits a newsletter subscription. Would you subscribe?

Biased Sampling Again

- It is natural to assume that the stocks in the emails are randomly chosen from a list of “buy” recommendations
- But suppose instead that different potential customers got different recommendations selected at random, and the recipients of “failed predictions” were then dropped from further notices
- If stock predictions are random (50% chance the stock will go up), then the odds of getting three hits in a row are 1 in 8
- That may be enough to attract lots of business!

Back to the Example

NEXNet is a relatively small but aggressive player in the telecommunications market in the mid-Atlantic region of the US. It is now considering a move into the Boston area.

NEXNet would like to *estimate* the average monthly phone bill in the communities of Weston, Wayland, and Sudbury, by conducting a phone survey. As an enticement for people to participate in the survey, NEXNet will offer discount coupons on certain products to survey participants.

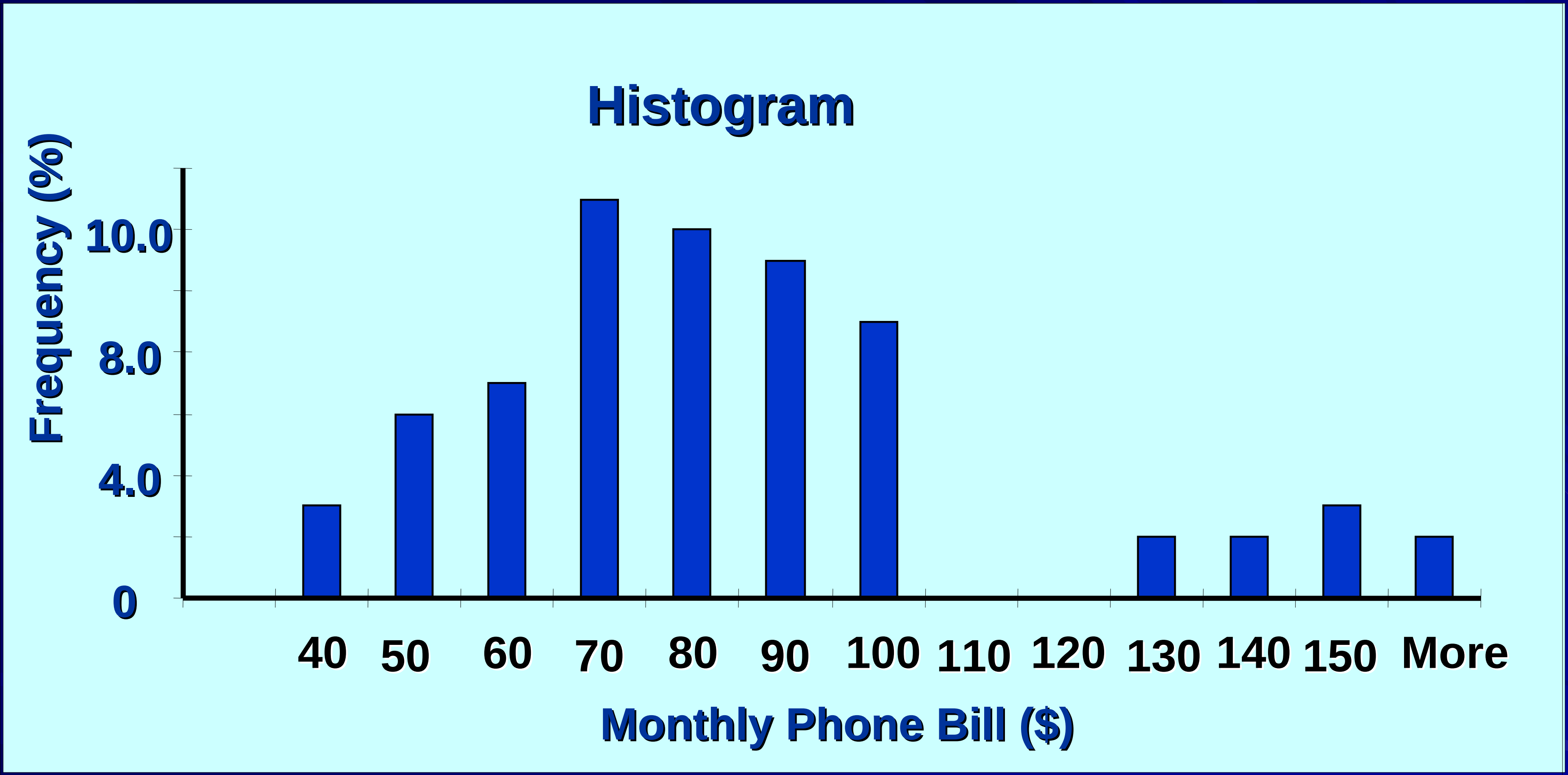
- How many households should NEXNet plan to survey (successfully) in order to effectively estimate the average phone bill in these three communities?

- How should NEXNet analyze the survey results?

Sample Histogram of May Phone Bills
(sample size n = 70)

Observation Number	May Phone Bill	Observation Number	May Phone Bill	Observation Number	May Phone Bill
1	$95.67	25	$79.32	49	$90.02
2	$82.69	26	$89.12	50	$61.06
3	$75.27	27	$63.12	51	$51.00
4	$145.20	28	$145.62	52	$97.71
5	$155.20	29	$37.53	53	$95.44
6	$80.53	30	$97.06	54	$31.89
7	$80.81	31	$86.33	55	$82.35
8	$60.93	32	$69.83	56	$60.20
9	$86.67	33	$77.26	57	$92.28
10	$56.31	34	$64.99	58	$120.89
11	$151.27	35	$57.78	59	$35.09
12	$96.93	36	$61.82	60	$69.53
13	$65.60	37	$74.07	61	$49.85
14	$53.43	38	$141.17	62	$42.33
15	$63.03	39	$48.57	63	$50.09
16	$139.45	40	$76.77	64	$62.69
17	$58.51	41	$78.78	65	$58.69
18	$81.22	42	$62.20	66	$127.82
19	$98.14	43	$80.78	67	$62.47
20	$79.75	44	$84.51	68	$79.25
21	$72.74	45	$93.38	69	$76.53
22	$75.99	46	$139.23	70	$74.13
23	$80.35	47	$48.06		
24	$49.42	48	$44.51		

THE HISTOGRAM



The Problem

- We will discuss how to determine the appropriate sample size n later.
- Our current problem is:

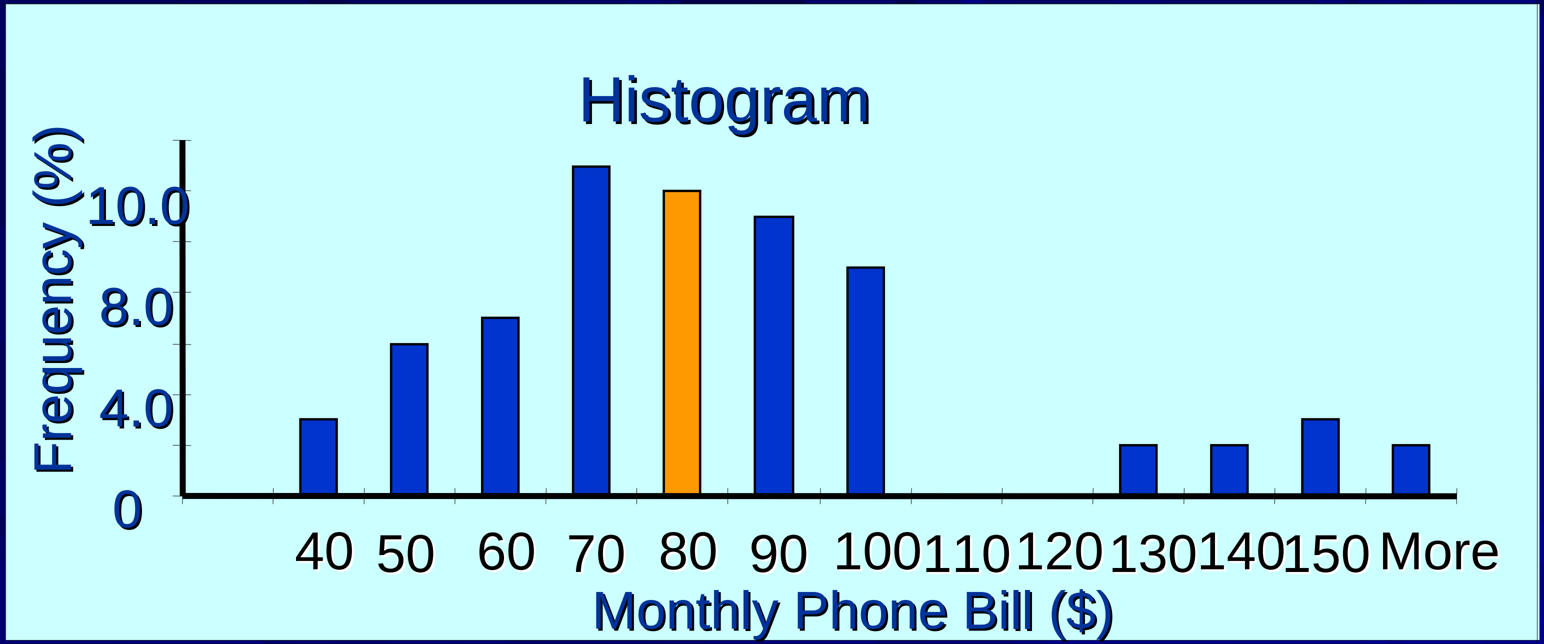
  Based on these n anticipated sample values $X_1, X_2, \ldots, X_n$, we want to make inferences about the entire population.
- Why? Because NEXNet has been profitable in communities with mean bills > $75, and no more than 15% households < $45 and at least 30% bills between $60 and $100

Estimates of the Population Mean

Sample Mean: sum of all the sample observations divided by the number of observations

$$\bar{X} = \frac{X_1 + X_2 + \ldots + X_n}{n}$$

Sample Median: the value that one-half the observations are below (50th percentile)



Sample Median = $76.65 Sample Mean = $79.40

- Sample mean accounts for the numerical value of each observation, but may be distorted by extreme values.

(This is the one we will use to estimate the population mean, μ.)

- Median is not affected by the magnitude of extreme values, but conveys information about position only.

Estimate of the Population Standard Deviation

- The sample variance s^2 is an "unbiased estimator" of the population variance, i.e., E [s^2] = Var[X] = σ^2.

- The sample standard deviation s is: $s = \sqrt{\dfrac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n - 1}}$

- We will use s to estimate the population standard deviation σ.

- Question: Why n - 1, and not n (as in the formula for calculating the population SD)?

- Answer: It gives a better (slightly larger) estimate. See:

http://mathcentral.uregina.ca/QQ/database/QQ.09.99/freeman2.html

- When n is large, the difference is negligible.

Example Continued

NEXNet arranged to have 70 randomly selected households successfully surveyed, as shown in the table. It found that the observed sample mean of the monthly phone bill was \$79.40, and the observed sample standard deviation was \$28.79.

- How would you characterize the shape of the distribution?
  Answer: It is not Normally distributed (some “outliers”).

- What is your estimate of the actual mean μ ?

$$\bar{x} = \$79.40$$

- What is your estimate of the actual standard deviation σ ?

$$s = \$28.79$$

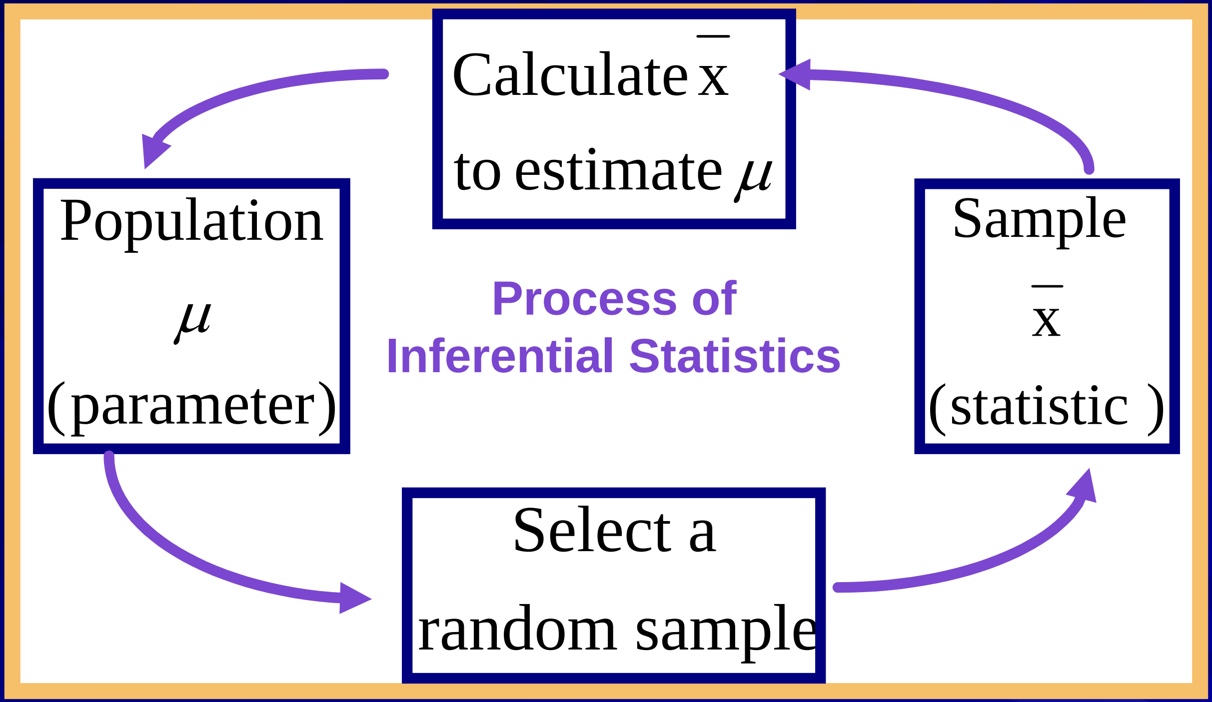
Clarify the Sampling Procedure

Before we collect the sample,

- $X_1, X_2, \ldots, X_n$ are the values that will arise from the sample
- $X_1, X_2, \ldots, X_n$ are random variables, i. i. d.
- As a result, we have for each X_i: $E[X_i] = \mu$, $Var[X_i] = \sigma^2$.
- $\overline{X} = \dfrac{X_1 + X_2 + \ldots + X_n}{n}$; The sample mean is a r.v. why?
- $s = \sqrt{\dfrac{\sum_{i=1}^{n} (X_i - \overline{X})^2}{n - 1}}$ is the sample standard deviation also a r.v. ?

Since both the sample mean and the sample standard deviation are r.v.'s, *we will get different results from different samples!*

Estimating the Population Mean Using the Sample Mean



- R.V. X (the population) : X represents a randomly selected item from the population.
- The sample mean $\bar{X}$ is also a R.V.

What is the variability of the mean?

Random variable $\bar{X}$ is defined as the average of n independent and identically distributed random variables, $X_1, X_2, \ldots, X_n$; with mean, μ, and Sd, σ. Then, for *large enough n* (typically $n \geq 30$), $\bar{X}$ is approximately Normally distributed with parameters: $\mu_{\bar{x}} = \mu$ and $\sigma_{\bar{x}} = \sigma/\sqrt{n}$

This result holds regardless of the shape of the X distribution (i.e. the Xs don't have to be normally distributed!)

And we can continue to estimate σ with s

Estimating the population mean using an Interval

- Idea: If we take a large enough random sample (i.e. n>=30) for r.v. X (i.e., the population of interest), then the sample mean, X, is approximately Normal and
- we can estimate the population mean, μ , using the interval shown.
- This interval denotes an area under the distribution of X which is +/- z standard deviations away from the mean.
- The value of z is determined by the “confidence level” assigned to the interval (see next slide), which depends on how much precision we need (or can afford)

Interval Estimate:

$$\overline{X} \pm Z \frac{\sigma}{\sqrt{n}}$$

or

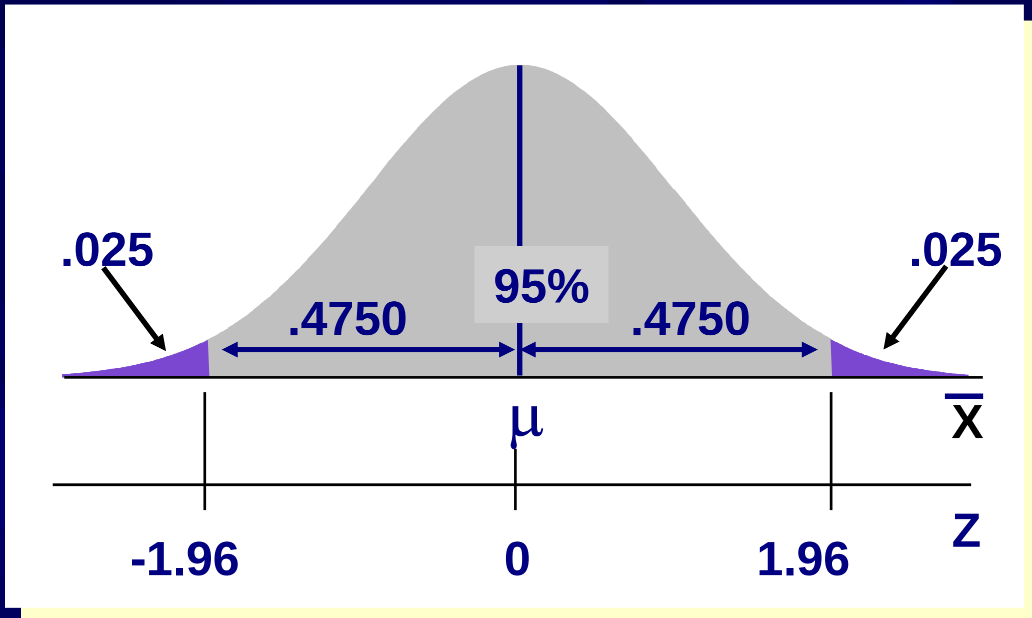
$$\overline{X} - Z \frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{X} + Z \frac{\sigma}{\sqrt{n}}$$

(In the interval above, if population SD, σ, is not known, use the sample SD:)

$$S^2 = \frac{\sum\left(X - \overline{X}\right)^2}{n-1}$$

$$S = \sqrt{S^2}$$

Values of Z for selected confidence levels:



Confidence Level	Z Value
90% (α=0.1)	1.645
95% (α=0.05)	1.96
98% (α=0.02)	2.33
99(% (α=0.01)	2.575

- We would, for example, say that we are 95% confident the true mean for x falls in the interval:

$$\bar{X} - 1.96\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}$$

(This means there is a .95 probability the interval given will contain the true mean.)

Example Continued

Calculate a 95% confidence interval for Nextel's mean monthly phone bill.

Formula:

$$\bar{X} \pm Z \frac{\sigma}{\sqrt{n}}$$

or

$$\bar{X} - Z \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + Z \frac{\sigma}{\sqrt{n}}$$

Data:

$\bar{X}$ = \$79.40; s = \$28.79; n=70;

For CL 95% z=1.96

1.96 * 28.79 / sqrt(70) = 6.74.

- We are 95% confident that the true mean μ is within 6.74 of the sample mean of 79.40 or [79.40 - 6.74, 79.40 + 6.74].

- The interval [72.66, 86.14] is called a 95% confidence interval (C.I.) for the population mean.

Example Continued

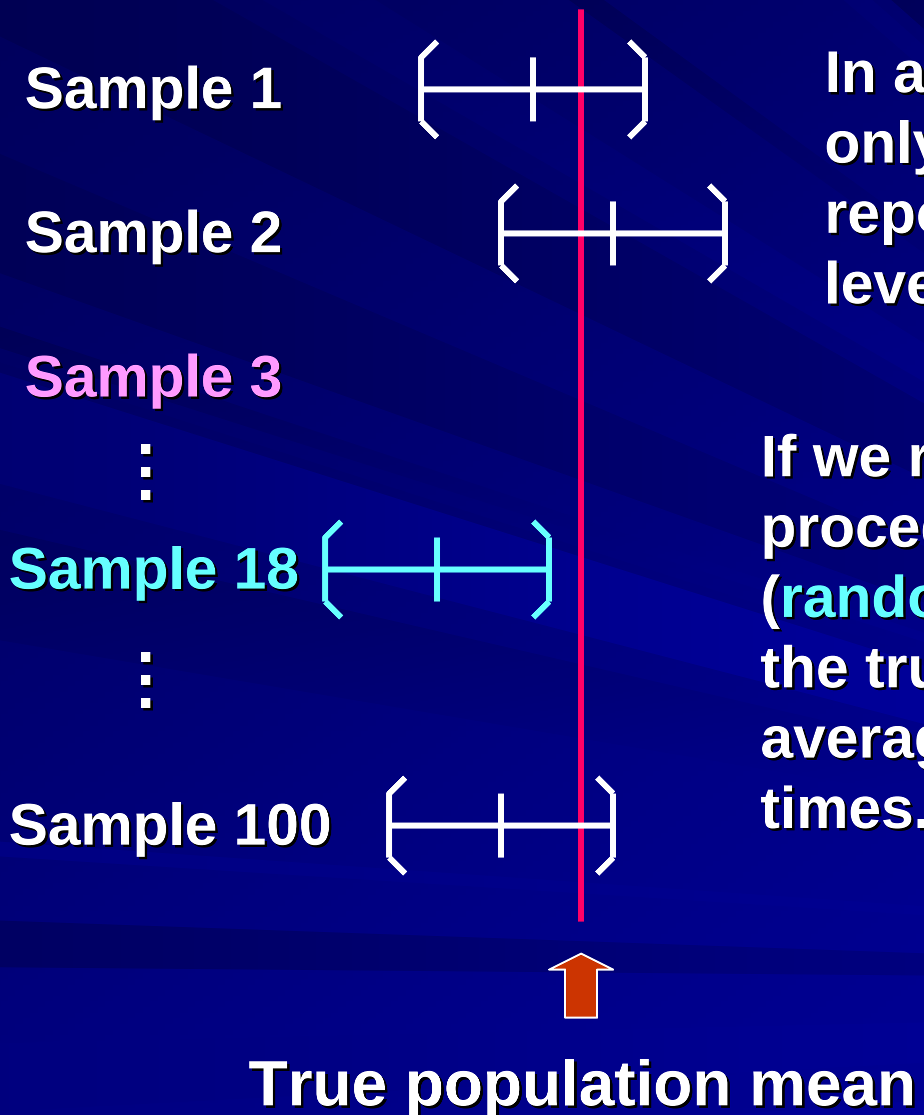
What if we want to be 99% confident ?

Use $z=2.575$

$2.58 * 28.79 / \mathrm{sqrt}(70) = 8.86.$

A 99% C.I. for μ is $[79.40 - 8.86, \; 79.40 + 8.86]$.

Interpreting confidence intervals



In a usual application, we only sample once and report a single confidence level, for example, 95%.

If we repeated this sampling procedure 100 times, our (random) intervals will capture the true population mean, on average, 95 times out of the 100 times.

An Example

- Each person take a coin and flip it ten times; count the number of heads and divide by ten
- This is your observed value of the proportion of heads
- Calculate the observed standard deviation s (heads=1, tails=0, use the formula for s)
- Calculate a 90% confidence interval for the proportion of heads from your individual data (z=1.65)
- We know the true (theoretical) mean is 5. Is the true mean outside your 90% confidence interval?
- Note that the true standard deviation is sqrt (n*p*[1-p]) = sqrt (2.5) = 1.58, so the 90% confidence interval is 2.39 to 7.61.

Interpreting confidence intervals

Sample 1

Sample 2

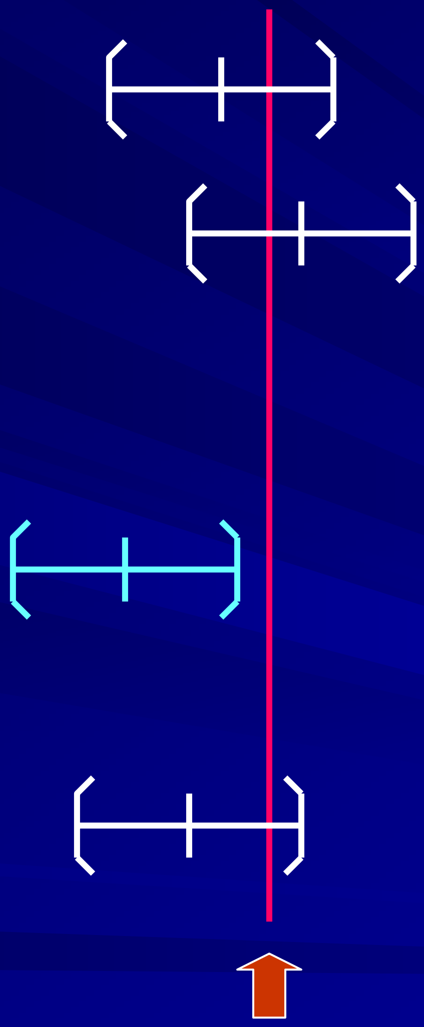
Sample 3

⋮

Sample 18

⋮

Sample 50



True population mean = .5

In a usual application, we only sample once and report a single confidence level, in our case, 90%.

When we repeat this sampling procedure 50 times, our (random) intervals will capture the true population mean, on average, 45 times out of 50.

Insights from the C.I. Formula

$$\bar{X} - Z\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + Z\frac{\sigma}{\sqrt{n}}$$

- Ideally, we want a tight interval with a high level of confidence (low α). But these are two conflicting goals!
- For a fixed sample size (n fixed), if we want to make a statement with a higher confidence level, we use a higher z which makes the interval wider: "The higher the confidence level the wider the interval."
- For a fixed confidence level (α and z fixed), if we increase the sample size n, then we get a narrower interval: "the larger the sample, the more accurate the estimate"
- For fixed sample size n and fixed confidence level, we can obtain a narrower interval if the population is less variable. "It is easier to make accurate inferences for populations with smaller SD"

Experimental Design: How large a sample do we need?

- Usually the goal is to reach an estimate of the mean which is within a certain tolerance value L from the population mean:

$$\bar{X} - L \le \mu \le \bar{X} + L$$

- From $\bar{X} - Z\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{X} + Z\frac{\sigma}{\sqrt{n}}$ we see that: $L = Z\frac{\sigma}{\sqrt{n}}$

- For a given z associated with a given CL, α, and given population SD, σ ,(or sample SD s). We can solve for the required sample size n (we always round up!)

$$n = \frac{z^2 s^2}{L^2}$$

Estimating Sample Size

- Suppose we needed to be 95% sure of being within \$4 of the true population mean, what sample do we need?

For confidence = 90 or α= 10,	z = 1.645
For confidence = 95 or α= 5,	z = 1.96
For confidence = 99 or α = 1,	z = 2.575

- L = 4, z = 1.96, and s = 28.79
- $n = z^2 s^2 / L^2 = 1.96*1.96*28.79*28.79/(4*4)$
- N = 199.01
- As a rule of thumb, n should always be rounded up to the nearest number, so we need a sample of 200

Another example: How large a sample do we need?

A marketing research firm wants to conduct a survey to estimate the average amount spent by each person visiting a popular resort. The survey planners would like to estimate the mean amount within (±) \$120, with 95% confidence.
(For the moment, assume that the population standard deviation of spending at the resort is σ = \$500.)

What is the sample size (n) you would need?

$$n = \frac{1.96^2 * 500^2}{120^2} = 66.69 \text{ (use } n=67)$$

If we don't know σ , we first estimate it with s in a pilot run.

Summary and Look Ahead

- Statistical sampling is about the value of information: how much information is needed, at what cost?
- Confidence intervals help us understand our level of uncertainty, which we can decide to reduce by collecting more data
- Next session we will talk about simulation, which helps us introduce uncertainty explicitly into our decision trees