Biostatistics and Informatics

BTM6000

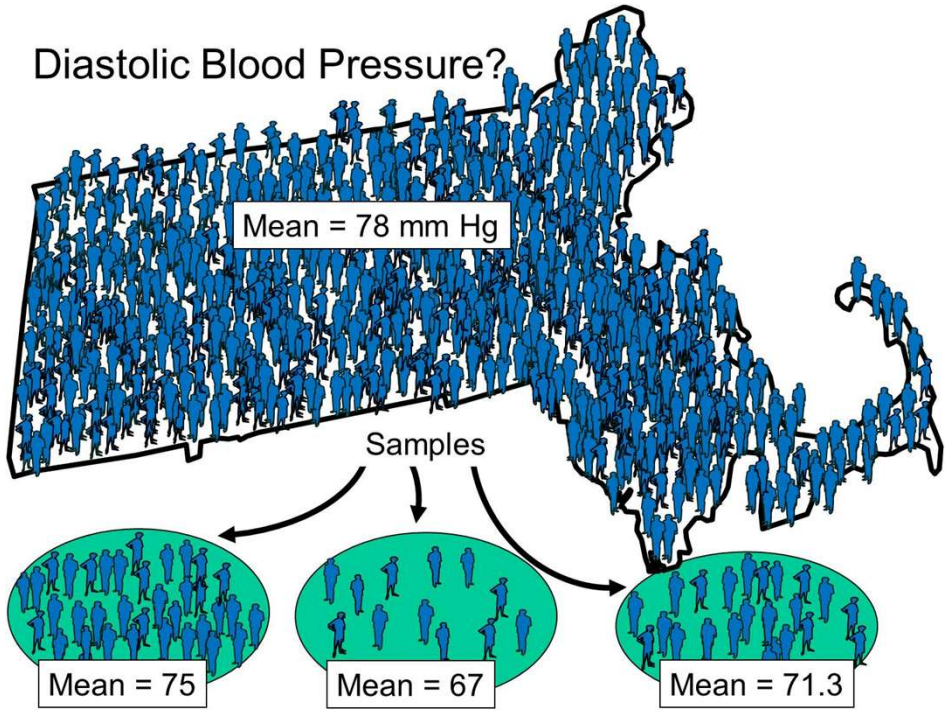
Session 4: Sampling Variability and the Central Limit Theorem

Class overview

- Population versus Sample
- Sampling Distributions, Central Limit Theorem
- Experimental vs. Observational Studies
- Prospective vs. Retrospective studies

- OpenIntro 4.1, 4.4 (sampling and CLT)
- OpenIntro 1.3-1.5 (study types)

Population versus sample

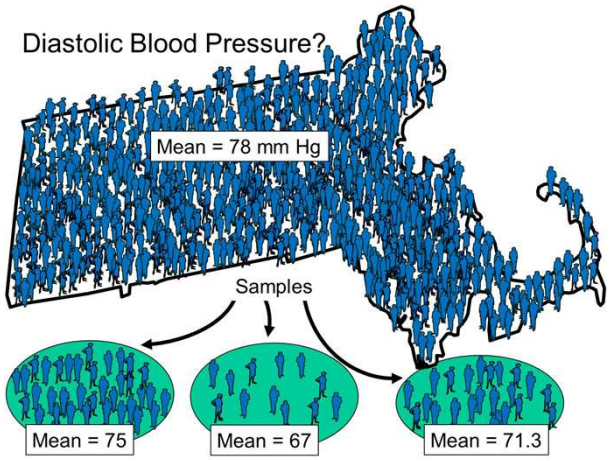


Population mean μ
(parameter)

inference

Sample mean $\overline{X}$

Sampling: vocabulary



Study units or *sampling units*: the individual elements of interest

Target population: the ideal population we would like to describe

Study population: the population from which we can actually sample. Systematic differences from target population is called *selection bias.*

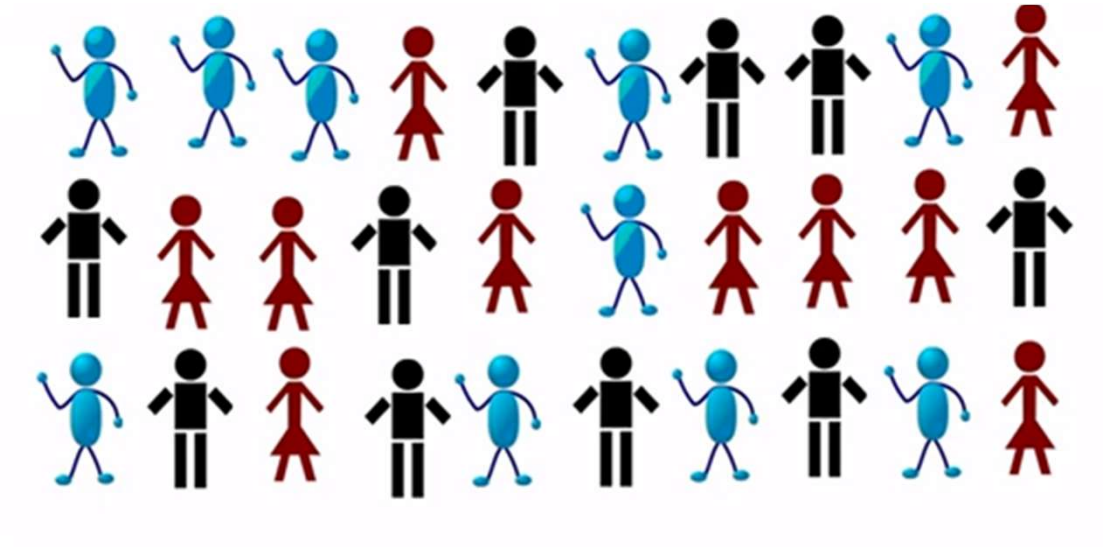
Sampling frame: a list of all units in the study population, used for sampling

Notation

- Greek letters for population *parameters*
  - Eg. population mean and standard deviation μ, σ
- English alphabet for sample *statistics*
  - Eg. sample mean and standard deviation $\overline{x}$, s

Sampling: strategies

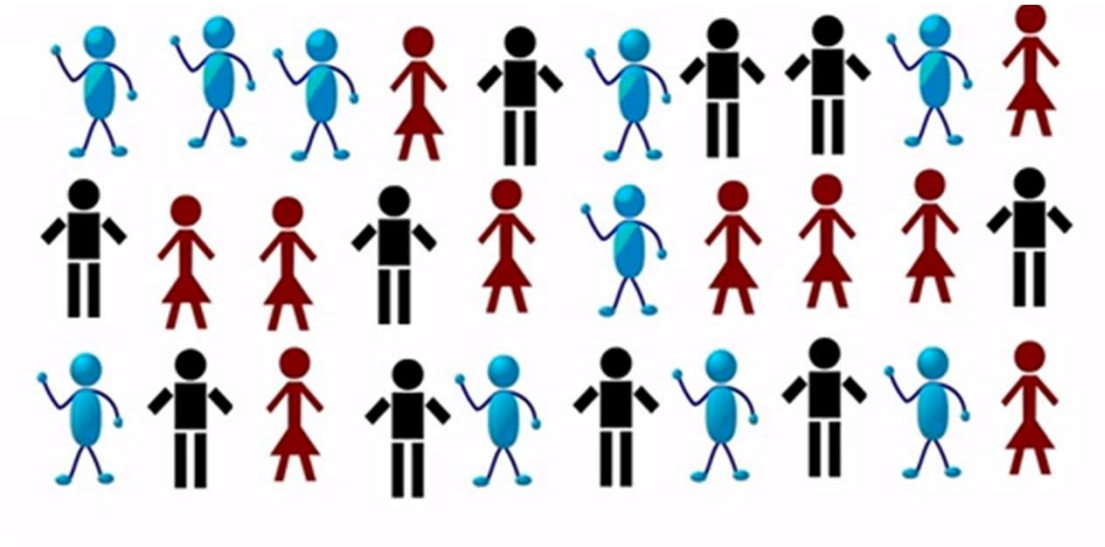
Simple Random Sample (SRS)



For n=10, sample(1:10, 10, replace=FALSE) #R code

Sampling: strategies

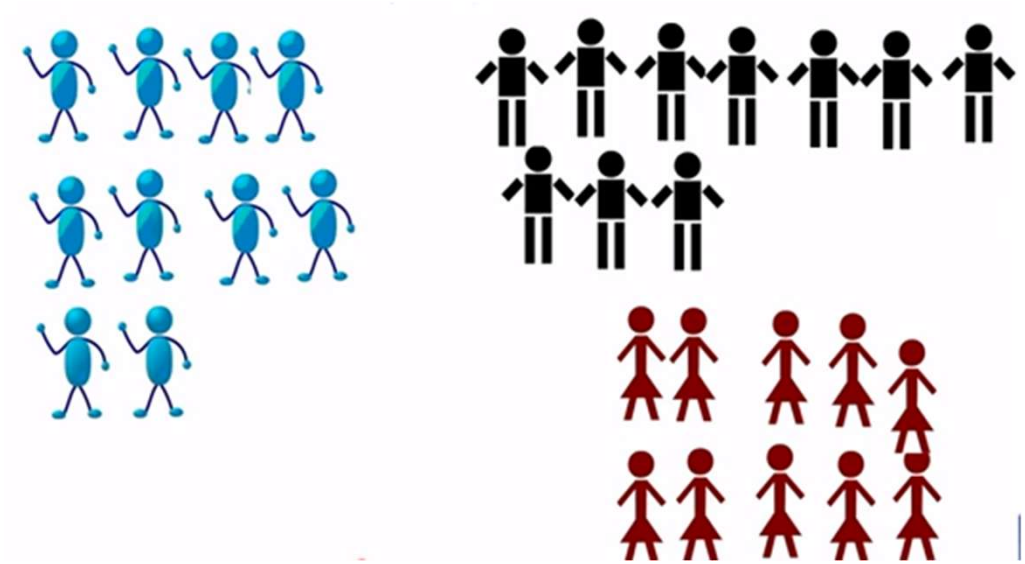
Systematic Sampling



For n=10, select a single number randomly between 1 and 30/10=3
e.g. 2, then 5, 8, …
Can often then treat as a SRS

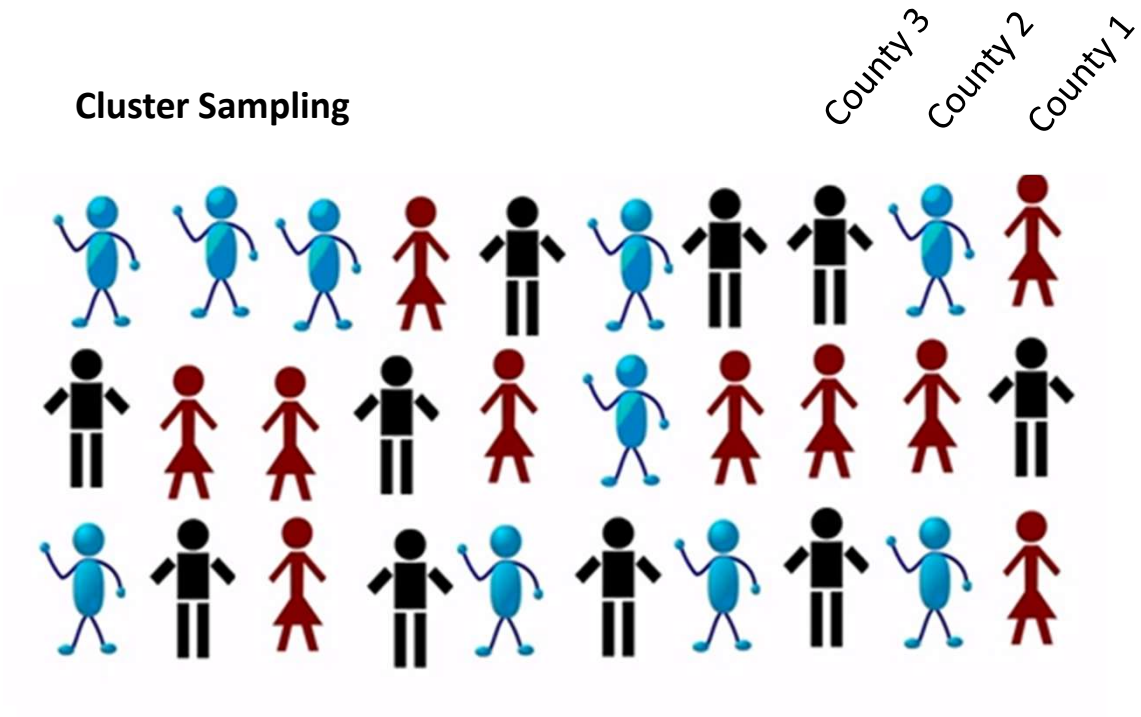
Sampling: strategies

Stratified Sampling



- Separate into *strata*
- Use SRS within each strata, with pre-defined n within each stratum
- If strata are well-chosen, sample variances are less than SRS

Sampling: strategies



Clusters are natural groupings of the units

- First take a SRS of clusters, then SRS within each cluster
- Sample variance is larger than SRS, depends on the number of clusters more than n selected from each cluster

Sampling: strategies

Nonprobability samples. E.g.:

- Volunteers
- Patients at a cancer clinic
- Other “convenience samples”

- Do not know the probability of selection
- Prone to selection bias, but often necessary

Sampling strategies example: geographic sampling

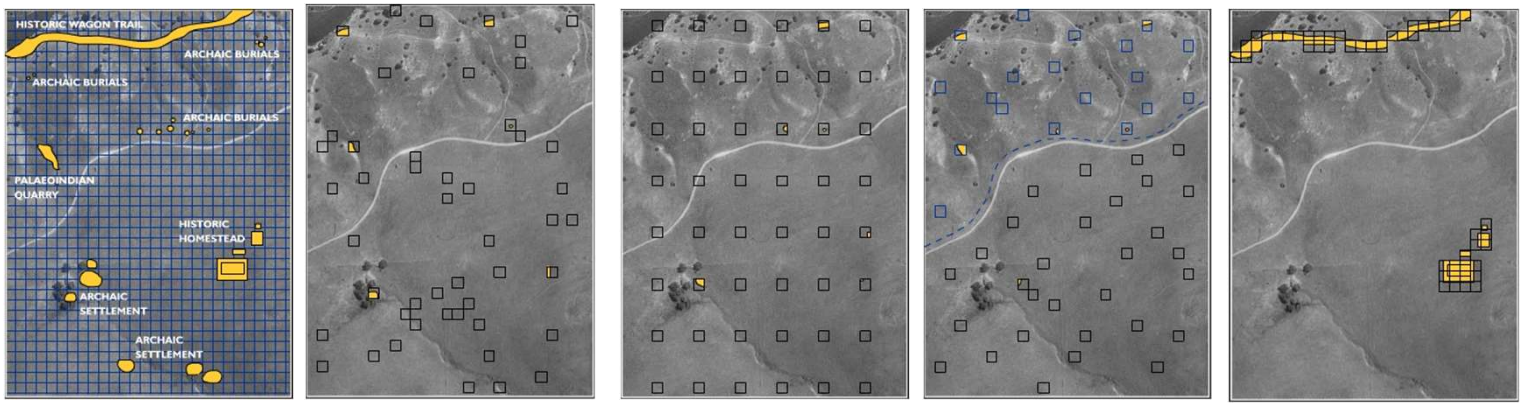
All sites

Simple Random Sampling (SRS)

Systematic Sample

Stratified Random Sample

Non-probability sampling



Experimental and Observational studies

Experimental – there is intervention.

- E.g. Randomized Controlled Trial (RCT)
- Normally there is randomized assignment to treatment and control groups
- Experimental studies are the only way to prove causality

Experimental and Observational studies

Observational – there is no intervention.

- **Cross-sectional**: a random sample of the population at a snapshot in time
  - e.g. NYC-HANES, BRFSS, NHANES
- **Retrospective**: subjects are recruited based on health status
- **Prospective**: a cohort is followed through time
  - E.g. Nurses' Health Study
- **Observational studies cannot prove causality**

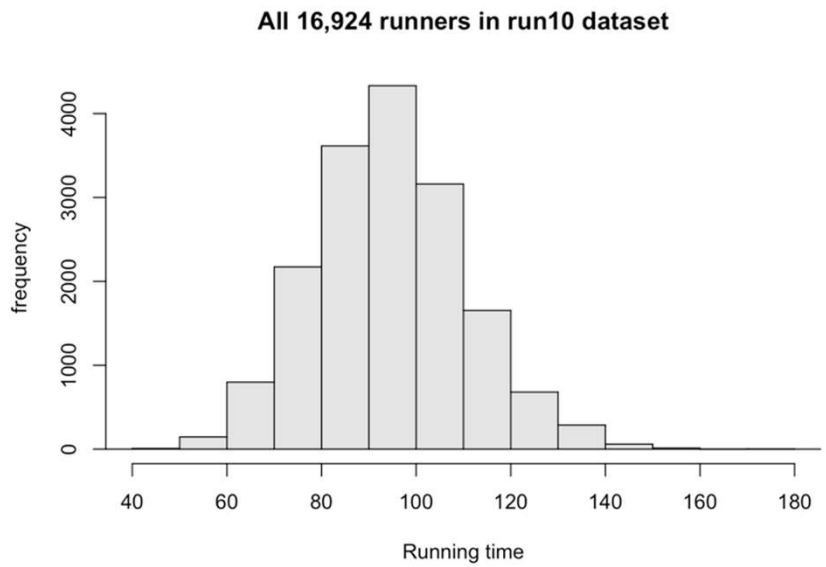
Experimental and Observational studies: does smoking cause lung cancer?

- Experimental approaches:
  1. Randomly assign never-smokers to smoking and non-smoking groups, follow for years to see who develops more lung cancer
  2. Randomly assign smokers to no-intervention or a cessation program, follow for years to see who develops more lung cancer

Experimental and Observational studies: does smoking cause lung cancer?

- Observational approaches:
  - **Cross-sectional**: Survey a sample of the population, ask whether they have ever been diagnosed with lung cancer and how much they have smoked
  - **Retrospective**: recruit 50 cases with lung cancer, and 50 controls who do not have lung cancer. See which group smokes more.
  - **Prospective**: observe groups of smokers and non-smokers for several years, see which group develops lung cancer more frequently

Population versus sample

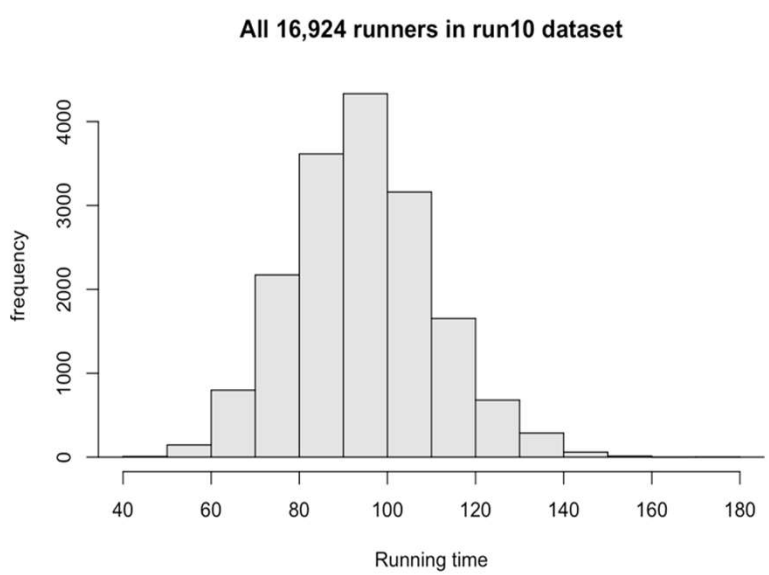


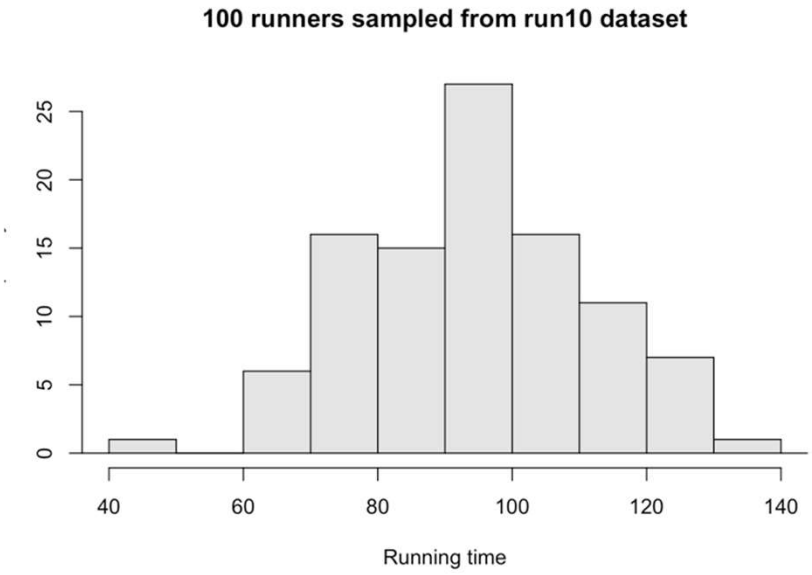
All 16,924 10 mile participants in the 2009 D.C. Cherry Blossom Run

Is this a sample or a population?

Population mean μ or sample mean $\overline{x}$?

Population versus sample



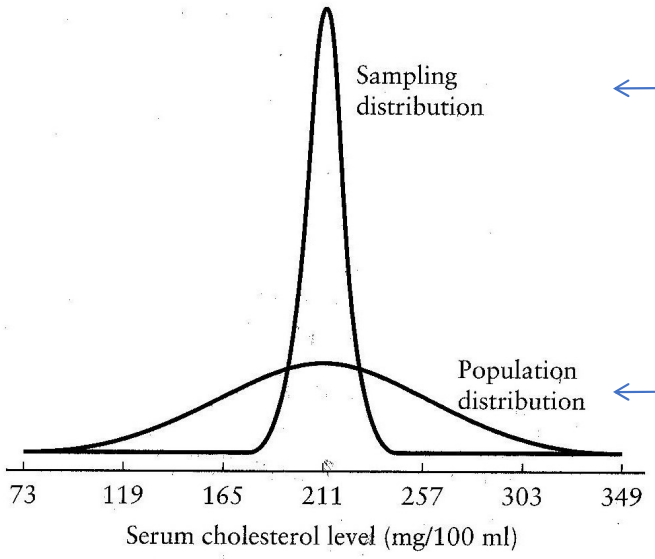


What is a “sampling distribution?”

It is the distribution of a statistic for many samples taken from one population.

1. Take a sample from a population
2. Calculate the sample statistic (e.g. mean)
3. Repeat. The distribution of the values obtained in (2) is a sampling distribution.

Sampling Distributions differ from Population Distributions

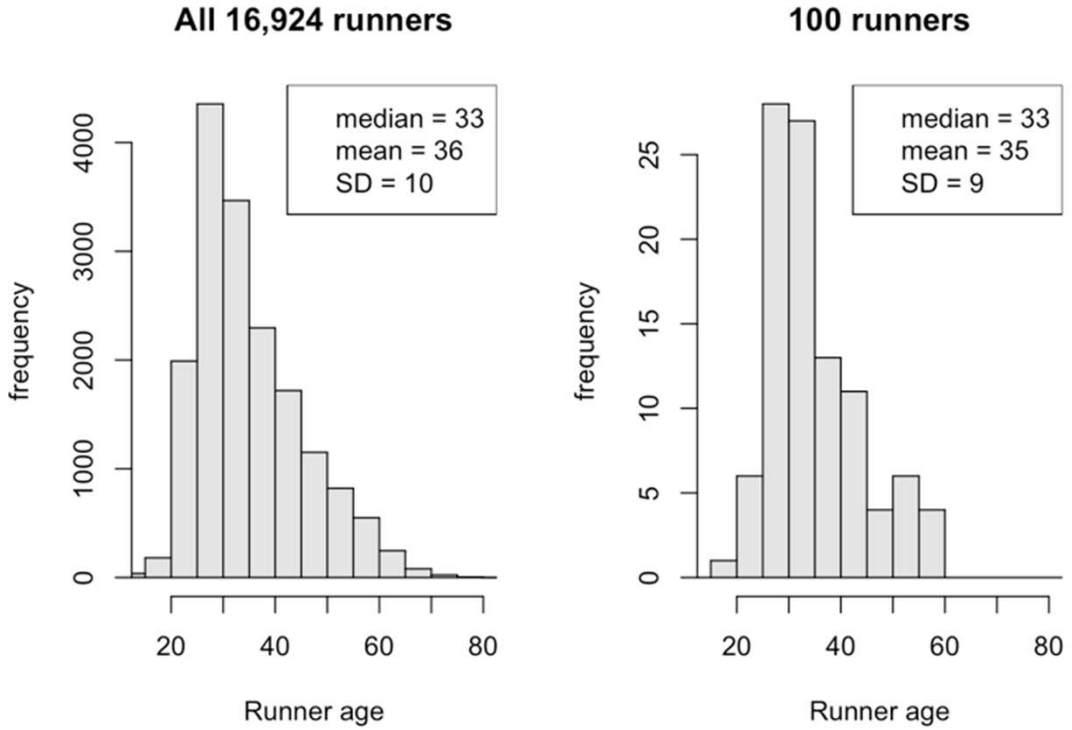


Mean serum cholesterol of groups (samples)

Serum cholesterol of individuals

FIGURE 8.1
Distributions of individual values and means of samples of size 25 for the serum cholesterol levels of 20- to 74-year-old males, United States, 1976–1980

Sampling Distributions differ from Population Distributions



Question: Is the histogram of a sample an approximation of the population distribution or the sampling distribution?

Answer: the population distribution!

Mean = 35 is only one point on the sampling distribution.

Central Limit Theorem

The "CLT" relates the sampling distribution (of means) to the population distribution.

1. Mean of the population (μ) and of the sampling *distribution* ($\overline{X}$) are identical

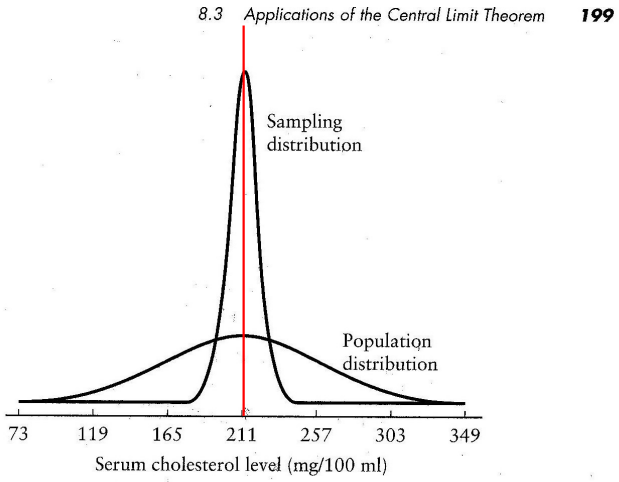


FIGURE 8.1
Distributions of individual values and means of samples of size 25 for the serum cholesterol levels of 20- to 74-year-old males, United States, 1976–1980

Central Limit Theorem

2. Standard deviation of the population (σ) is related to the standard deviation of the distribution of sample means by:

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

σ = standard deviation of the population
SE = standard error of the mean
n = sample size

sampling distribution spread is SE

Population spread is σ

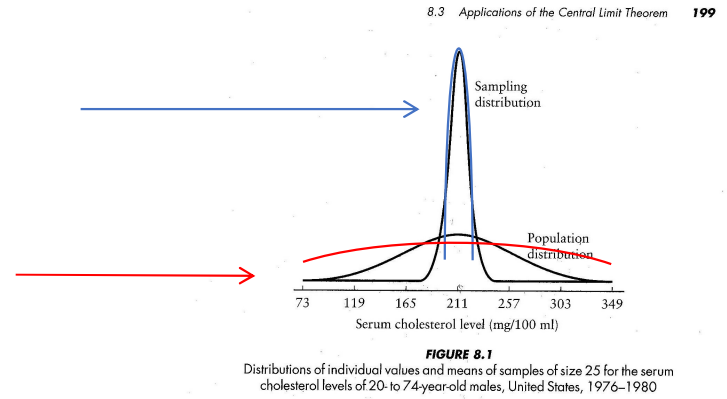
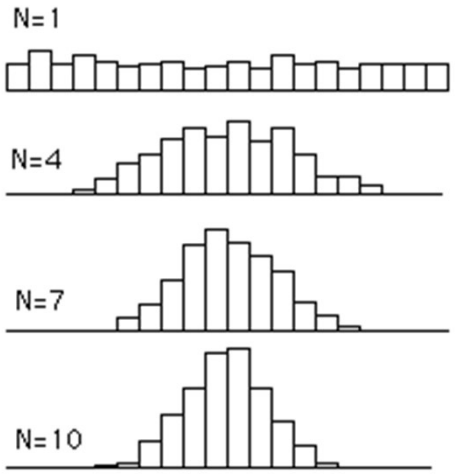
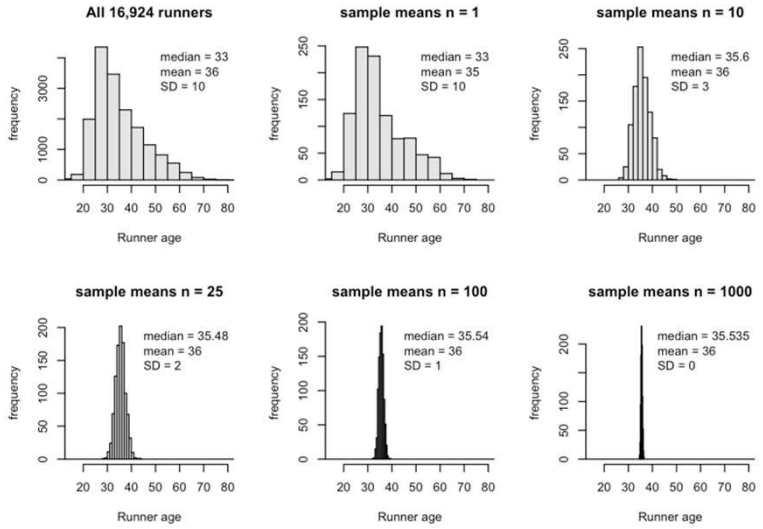


FIGURE 8.1
Distributions of individual values and means of samples of size 25 for the serum cholesterol levels of 20- to 74-year-old males, United States, 1976–1980

Central Limit Theorem

3. For large n, the shape of the sampling distribution of means becomes normal

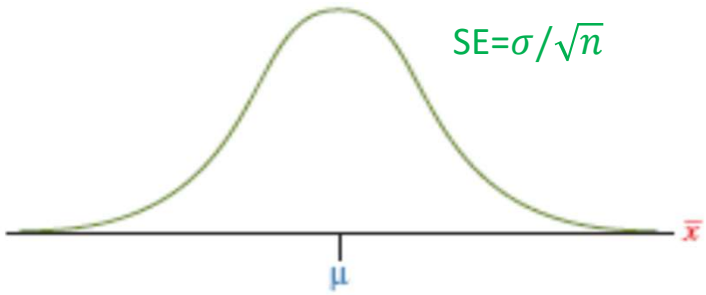




run10 dataset

The Sampling Distribution of the Sample Mean

- For a random sample of size n from a population having mean μ and standard deviation σ, the sampling distribution of the sampling mean has the following distribution:



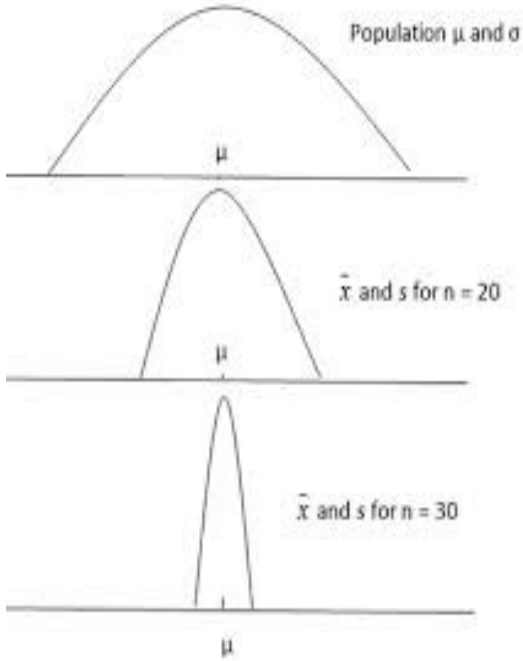
Center is μ
"Standard Error" or SE = σ/√n

Note: standard error is just a standard deviation, but we use this name to indicate that we are looking at the sampling distribution, not an ordinary distribution

Standard Error

- The standard error tells us how much variability to expect in sample statistics “just by chance”
- Sampling variability decreases

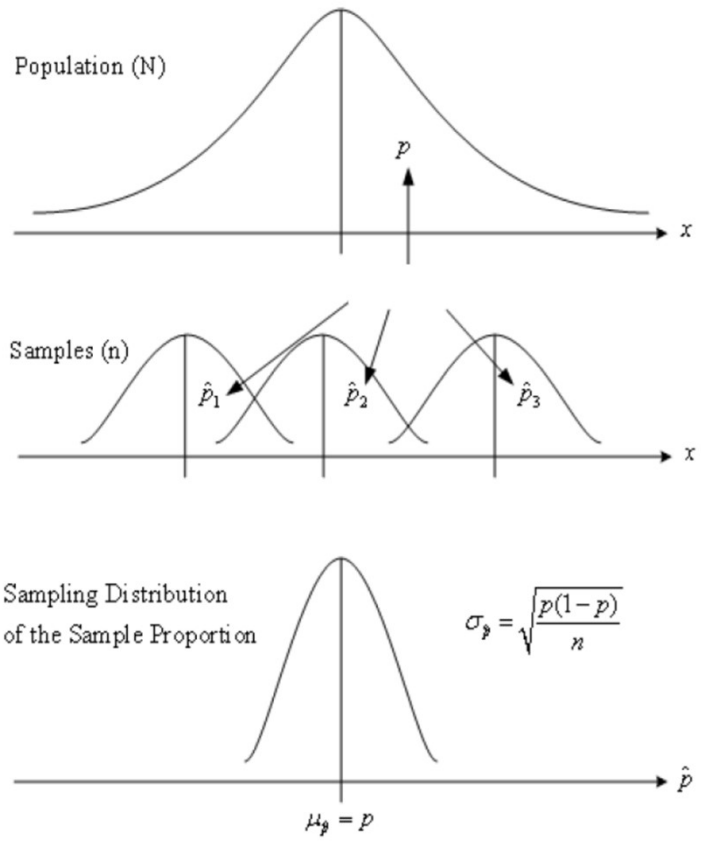
  with larger sample sizes
  - The sample mean will tend to fall closer to the population mean



CLT: How Large a Sample?

- The sampling distribution of the sample mean takes **more of a bell shape** as the random sample size *n* increases.
  - If the population distribution is skewed then a larger *n* is needed before the shape of the sampling distribution is close to normal.
    - In practice, the sampling distribution is usually close to normal when the sample size *n* is at least about 30.
  - If the population distribution is approximately normal, then the sampling distribution is approximately normal for *all* sample sizes.

Sampling Distribution of a Proportion



For a random sample of size *n (large)* from a population with proportion *p* of outcomes in a particular category, the sampling distribution is:

-normally distributed

-centered at the population proportion p

-with standard error of:

$$SE_p = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Applications of the Central Limit Theorem: Inference

For sufficiently large n,

$$Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$$

is normally distributed with mean 0 and standard deviation 1.

- With σ known or estimated, can calculate Z relative to *sampling distribution* to infer whether a sample came from a theoretical population of mean μ.

Applications of the Central Limit Theorem: Confidence Intervals

$$Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$$

Can estimate 95% confidence interval on the estimate of $\bar{X}$, given assumptions about μ, σ:

$$-1.96 \leq Z \leq 1.96$$

Central Limit Theorem: Summary

- Forms the basis for the t-test, Analysis of Variance (ANOVA), and confidence intervals
- Tells us the sampling distribution for large enough n, even if we do not know the population distribution
  - "large enough" depends on how normal the population distribution is
  - The size of the sample (n) matters, the size of the population *does not matter* if you have a representative sample

Summary

- What is a sampling distribution, and how does it relate to the population distribution?
  - Sampling distribution of means is related to population distribution by the Central Limit Theorem
- Some general classes of studies:
  - Experimental studies
  - Observational studies
    - retrospective and prospective studies