STOR 455
STATISTICAL METHODS I

Jan Hannig

Fundamental Concepts

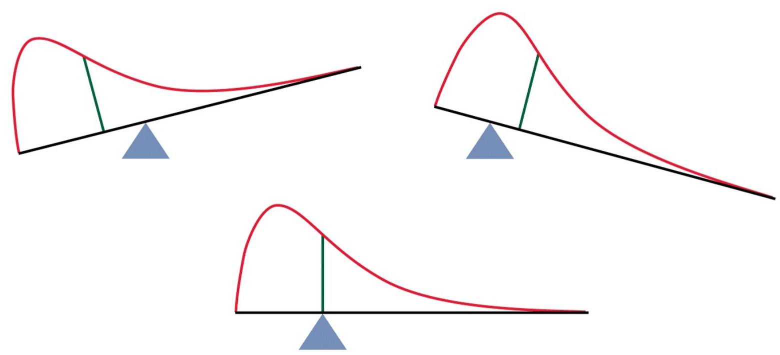
- Population: the entire group of individuals that we want information about.
- Sample: a part of the population that we actually examine in order to gather information.
- Statistical inference: to make an inference about a population based on the information contained in a sample.
- A *model* is mathematical description of the quantities of interest
  - Gaussian with unknown mean and variance
- A *parameter* is a value that describes the population. It's fixed but unknown in practice.
  - the mean and variance of the SAT score of all the students, who are about to take it.
- A *statistic* is a value that describes a sample. It's known once a sample is obtained.
  - The mean and variance SAT score of all the students, who are selected into the study.
  - A sample analogy of the parameter.

Inferential Procedures

- Once the sample is selected, inference about the sample is performed
- Types of inference
  - Point Estimation
  - Confidence Intervals
  - Hypothesis Testing

Parameter

- Mean
  - If the population consists of equally likely numbers ($y_1,\ldots,y_N$) then $\mu = \frac{1}{N}\sum_{i=1}^{N} Y_i$ [Notice the upper case]
  - If not equally likely $\mu = \sum_{i=1}^{N} Y_i p_i$ where p_i is the probability. (Explain on a picture)



Point Estimate

- Our sample consists of *n* randomly chosen observations $(y_1,\ldots,y_n)$.
- Based on this sample we estimate the population mean μ by sample mean

$$\hat{\mu} = \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

[Notice the LOWER case]

Parameter

- Standard Deviation
  - If the population consists of equally likely numbers ($Y_1,\ldots,Y_N$) then

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \mu)^2}$$

  - If not equally likely

$$\sigma = \sqrt{\sum_{i=1}^{N}(Y_i - \mu)^2 p_i}$$

- Variance = (standard deviation)^2

Point Estimate

- Our sample consists of *n* randomly chosen observations ($y_1,\ldots,y_n$).
- Based on this sample we estimate the population sd σ by sample sd

$$\hat{\sigma} = \sqrt{\frac{SSY}{n-1}}, \quad SSY = \sum_{i=1}^{n}(y_i - \bar{y})^2$$

  – SSY stands for "Sum of Squares for Y"

Parameter

- Correlation
  - If items are equally likely

$$\rho_{Y,X} = \frac{\sum_{I=1}^{N} (Y_I - \mu_Y)(X_I - \mu_X)}{\sqrt{\left[\sum_{I=1}^{N} (Y_I - \mu_Y)^2\right]\left[\sum_{I=1}^{N} (X_I - \mu_X)^2\right]}}$$

  - Items are not equally likely

$$\rho_{Y,X} = \frac{\sum_{i=1}^{N} (Y_i - \mu_Y)(X_i - \mu_X) p_i}{\sigma_X \sigma_Y}$$

- Meaning
  - Always $-1 \le \rho \le 1$, if independent $\rho = 0$

Point Estimate

- Our sample consists of *n* randomly chosen pairs of observations $((x_1, y_1), \ldots, (x_n, y_n))$.
- Based on this sample we estimate the population correlation ρ by sample correlation

$$\hat{\rho} = r = \frac{SXY}{\sqrt{SSX \cdot SSY}}$$

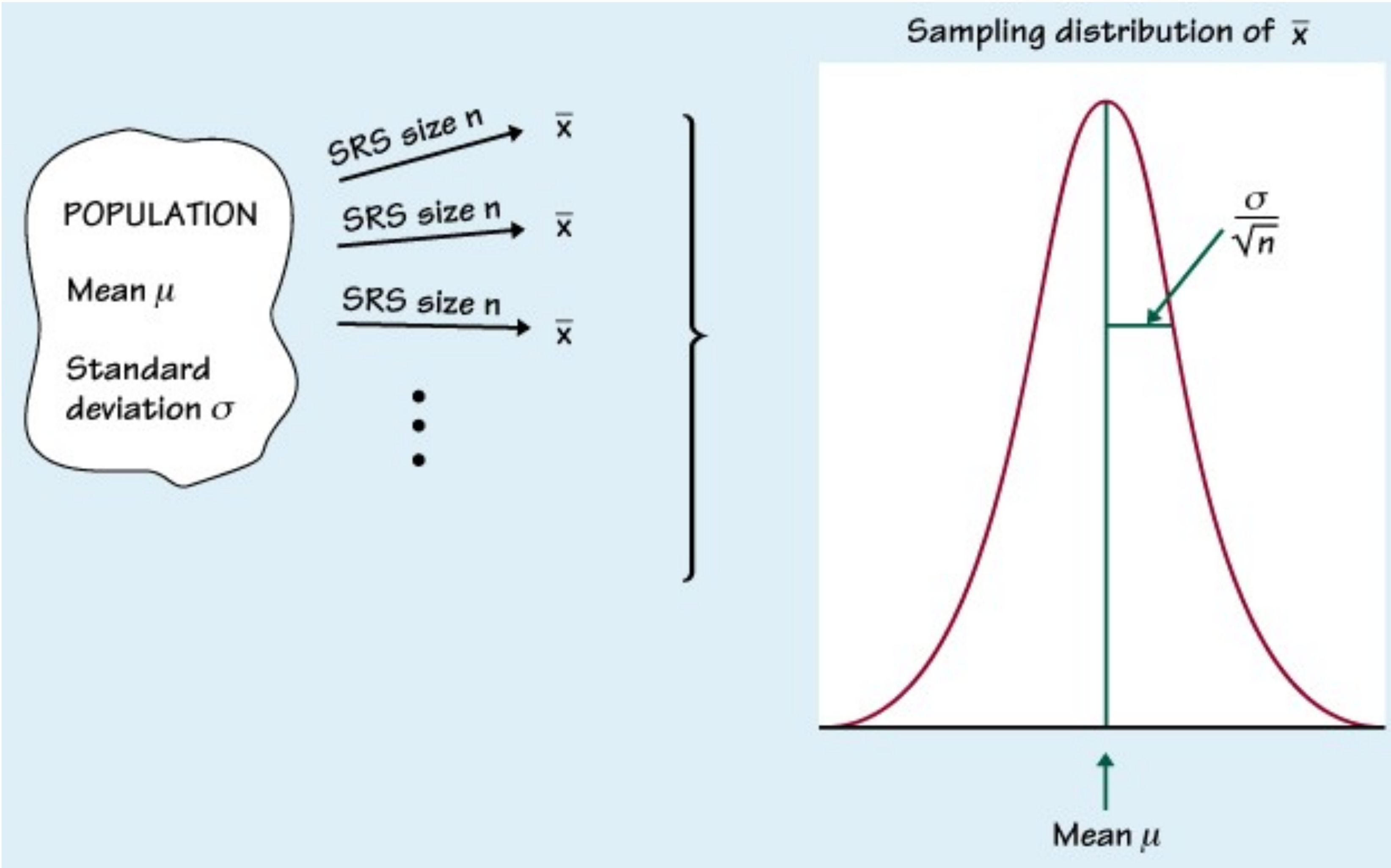
$$SXY = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}), \quad SSX = \sum_{i=1}^{n} (x_i - \bar{x})^2$$

  – SXY stands for "Sum for XY"

Unbiased Estimate

- If a *different sample* of the same sample size was selected from the same population, a slightly *different value* of an estimator would be obtained.
- An estimator is unbiased if the *average* of the estimator computed over *all possible samples* is equal to the *parameter value*.
- Example - finite population

Example


Sampling distribution of $\bar{x}$
POPULATION
Mean μ
Standard deviation σ
SRS size n
SRS size n
SRS size n
$\bar{x}$
$\bar{x}$
$\bar{x}$
$\frac{\sigma}{\sqrt{n}}$
Mean μ

Sampling Distribution of $\overline{X}$

SAMPLING DISTRIBUTION OF A SAMPLE MEAN

If a population has the $N(\mu, \sigma)$ distribution, then the sample mean $\overline{x}$ of n independent observations has the $N(\mu, \sigma/\sqrt{n})$ distribution.

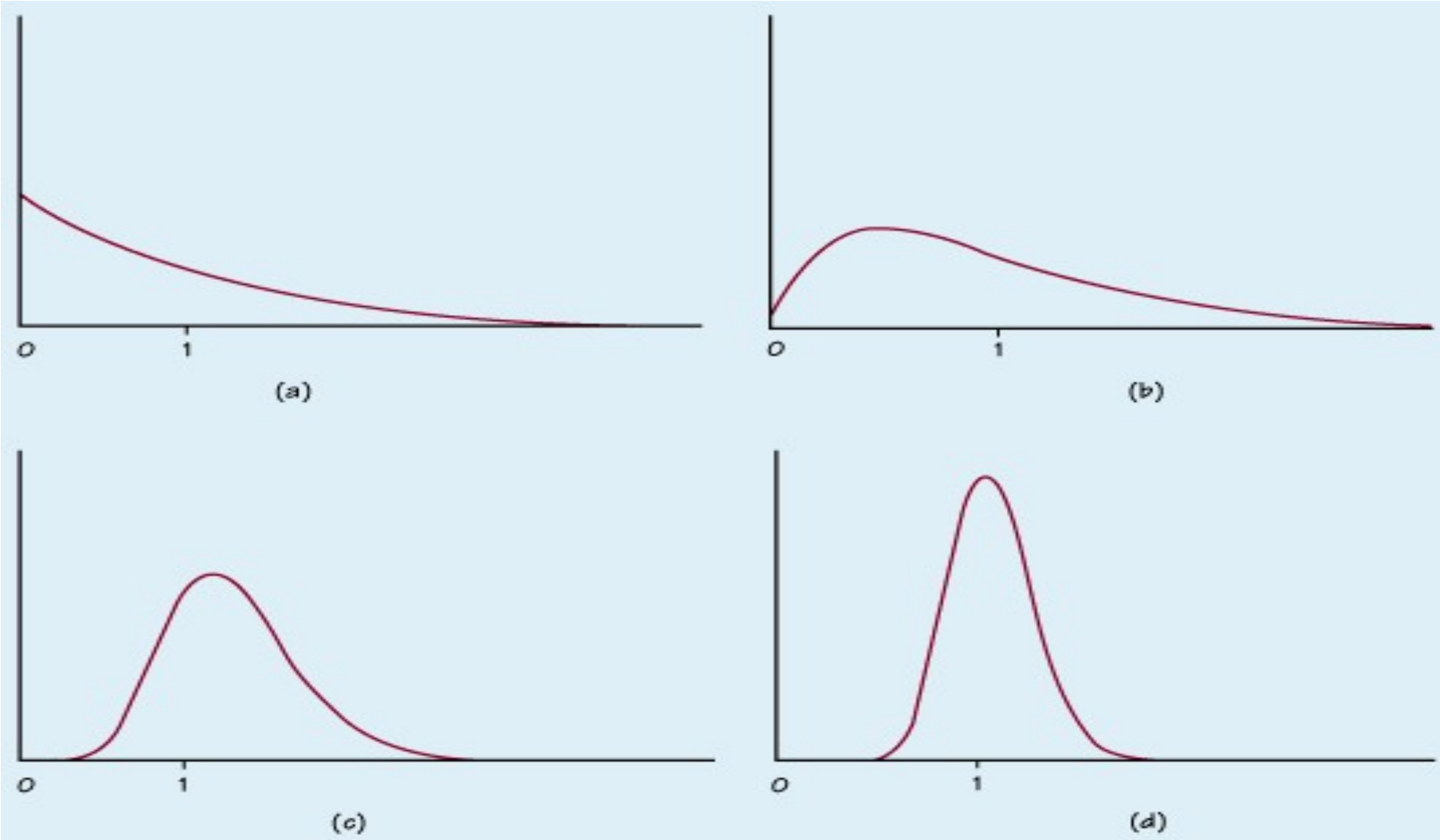
Definition, pg 362a
Introduction to the Practice of Statistics, Fifth Edition
© 2005 W. H. Freeman and Company

CENTRAL LIMIT THEOREM

Draw an SRS of size n from any population with mean μ and finite standard deviation σ. When n is large, the sampling distribution of the sample mean $\overline{x}$ is approximately normal:

$$\overline{x} \text{ is approximately } N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

Definition, pg 362b
Introduction to the Practice of Statistics, Fifth Edition
© 2005 W. H. Freeman and Company



The distributions of $\bar{X}$ for

(a). 1 obs. (b). 2 obs. (c). 10 obs. (d). 25 obs.

Point Estimation

- A point estimator draws inference about a population by estimating the value of an unknown parameter using a single value or a point.
- For example, sample mean estimates pop. Mean.
- Drawbacks:
  - How different is the estimate from the true parameter?
  - How reliable is our estimate?
  - How confident are we with our estimate?
  - Ways to improve?

Confidence Interval

- A confidence interval has (usually) the form:

  point estimate ± margin of error

- Symmetric about the point estimate (PE)
  - The PE is our guess for the value of the unknown parameter.
  - The margin of error (ME) shows how accurate we believe our guess is, based on the sampling distribution of the estimate.

Confidence Interval

- *1-α*: confidence level,
  - how confident we are that the confidence interval will cover the true population mean.
- Want to find a level C=*1-α* confidence interval for θ
- Such that P($L \leq \theta \leq U$)=1-α.
  - The meaning of this is "repeated sampling" probability [see (1.6.5) in the book]
  - In many applications we have

$$\hat{\theta} - \text{table value} \cdot SE(\hat{\theta}) \leq \theta \leq \hat{\theta} + \text{table value} \cdot SE(\hat{\theta})$$

  - *SE* is the *standard error (estimate of the sd)* of the point estimator.