
General Biostatistics

Part 3

1

Methods for Statistical Inference

Estimation

Marie Diener-West, Ph.D.
Department of Biostatistics
Johns Hopkins University
Bloomberg School of Public Health

Outline

- Sampling distributions
- Estimation
 - Point estimation
 - Confidence interval estimation
- 95% confidence interval
- Z versus t distributions
- Example
- Estimation and sample size

3

Sampling Distributions

- A sampling distribution is the *theoretical* distribution of all possible values which can be assumed by some sample statistic computed from samples of the same size randomly drawn from the same population.

4

Simple Random Sample

- A **simple random sample** is a sample of size n drawn from a population of size N in such a way that every possible sample of size n has the same probability of selection.
- Practical sampling methods:
 - Picking numbers from a hat.
 - Random number table or generator, computer programs

5

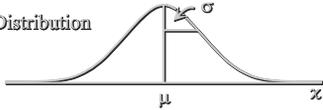
Three Distinct Distributions

- The population distribution describes the shape, central tendency and spread of the N population values
- The distribution of the sample describes the shape, central tendency and spread of the n sampled values
- The sampling distribution describes the shape, central tendency and spread of the summary statistics of all possible samples of size n taken from the population

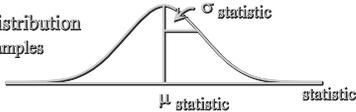
6

Sampling Distribution

Population Distribution



Sampling Distribution
All possible samples
of size n



Useful Sampling Distributions

- The most frequently used sampling distributions describe
 - the sample mean
 - the sample proportion
 - the difference between two sample means of samples from two different populations
 - the difference between two sample proportions of samples from two different populations

Sampling Distribution of the Mean

- What is the sampling distribution of the sample mean?
 - The statistic of interest is the sample mean
 - The sampling distribution describes the probability associated with every possible sample mean
 - Some observed sample means have high probability, some have low probability

Sampling Distribution of the Mean

- Example: A "population" of 10 student ages ($N=10$)
 - 27, 28, 31, 35, 35, 40, 42, 43, 50, 52
 - Remember $\mu = 38.3$ years and $\sigma = 8.6$ years
- Take one sample of size $n = 4$
 - Possible sample means?
 - The lowest possible sample mean is 30.25
 - The highest possible sample mean is 46.75
 - The most "typical" sample mean is 38.3

10

Sampling Distribution of the Mean

- Population Distribution: 10 student ages
 - 68% of ages fall between $\pm \sigma$ years of the true mean
 - 95% of ages fall between $\pm 2\sigma$ years of the true mean
 - 99% of ages fall between $\pm 3\sigma$ years of the true mean
- Sampling distribution of size $n = 4$
 - 68% of mean ages fall between $\pm \sigma/\sqrt{n}$ years of the true mean
 - 95% of mean ages fall between $\pm 2\sigma/\sqrt{n}$ years of the true mean
 - 99% of mean ages fall between $\pm 3\sigma/\sqrt{n}$ years of the true mean

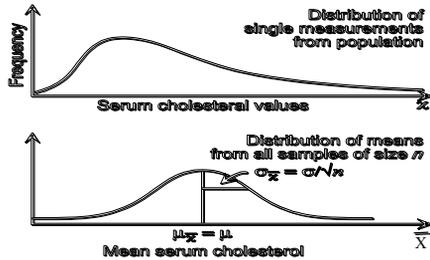
11

Central Limit Theorem

- Given a population of any non-normal continuous distribution with mean μ and standard deviation σ , the sampling distribution of \bar{X} will be approximately normally distributed with mean μ and standard deviation σ/\sqrt{n} , when the sample size is large.

12

Central Limit Theorem



13

Inferential Methods

- Estimation
 - Point estimation
 - A sample statistic is an **estimator** of a population parameter; the value of the statistic for a particular sample is an **estimate**.
 - Interval estimation
 - A point estimate surrounded by an interval that expresses the uncertainty or variability associated with the estimate.
- Hypothesis testing (next class)

14

Estimation

- Point estimate
 - The best estimate of the true population parameter based on the data.
- Interval estimate
 - Confidence interval

15

Point Estimation

- Single sample
 - The sample mean is the “best” estimate of the population mean
 - The sample proportion is the “best” estimate of the population proportion

16

Point Estimation

- Two independent samples
 - The difference between 2 sample means (from 2 samples taken from 2 independent populations) is the “best” estimate of the true difference in the 2 population means
 - The difference between 2 sample proportions (from 2 samples taken from 2 independent populations) is the “best” estimate of the true difference in the 2 population proportions

17

Confidence Interval Estimation

- A 95 % confidence interval (CI) for a true but unknown population parameter
 - sample statistic $\pm z_{0.05/2}$ (standard error)
- Interpretations
 - In repeated sampling from a normally distributed population, 95% of intervals will include the true population parameter
 - “We are 95% confident that the interval contains the true population parameter.”

18

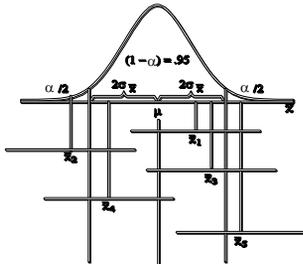
95% CI for Population Mean

- A 95% confidence interval for the true population mean is based on the sample mean
- sample statistic $\pm Z_{0.05/2}$ (standard error)

$$\bar{x} \pm 1.96 \frac{\sigma}{\sqrt{n}}$$

19

95% Confidence Intervals



20

The t Distribution

- Alternative distribution when the true population standard deviation, σ , is unknown
- Mean = median = mode = 0
- Symmetrical about the mean
- t ranges from $-\infty$ to $+\infty$
- Family of distributions, determined by $n-1$, the degrees of freedom
- t approaches Z as n increases

21

Z versus t

- When the true population standard deviation, σ , is unknown then use the sample standard deviation, s , as the best available estimate
- sample statistic $\pm t_{0.05/2, df}$ (standard error)

$$\bar{x} \pm t_{0.05/2, df = n-1} \frac{s}{\sqrt{n}}$$

22

Example: Student Ages

- Example: 10 student ages (N=10)
 - 27, 28, 31, 35, 35, 40, 42, 43, 50, 52
 - Remember $\mu = 38.3$ years and $\sigma = 8.6$ years
- Take a sample of size $n = 4$
 - Sample: 28, 42, 43, 50
 - We can calculate

$$\bar{x} = 40.75, \quad s = 9.2$$

23

Example: Student Ages

- 95% confidence interval for the true population mean $\bar{x} \pm t_{0.05/2, 3 \text{ df}} \frac{s}{\sqrt{n}}$
- $t_{0.05/2, 3 \text{ df}}$ is based on $\alpha=0.05$ (2-sided) and $df = n-1 = 3$
- $t = 3.182$ from the table

24

Example: Student Ages

- 95% confidence interval (CI):
 $40.75 \pm 3.182 (9.2/\sqrt{4}) = 40.75 \pm 14.6 = (26.1, 55.4)$
- “I am 95% confident that this interval contains the true population mean.”

25

Example: Student Ages

- What happens to the 95% confidence interval for the true population mean as the sample size increases?

$$\bar{x} \pm t_{0.05/2, 3 \text{ df}} \frac{s}{\sqrt{n}}$$

- The confidence interval becomes narrower.

26

Another Example: Birth Weights

- Take a sample of $n=25$ infants
- We calculate
 $\bar{x} = 2500 \text{ gm}, s = 900 \text{ gm}$
- 95% confidence interval
 $\bar{x} \pm t_{0.05/2, 24 \text{ df}} \frac{s}{\sqrt{n}}$
- $t_{0.05/2, 24 \text{ df}}$ is based on $\alpha=0.05$ (2-sided) and $\text{df} = n-1 = 24$
- $t = 2.064$ from the table

27

Another Example: Birth Weights

- 95% confidence interval (CI):
 $2500 \pm 2.064 (900/\sqrt{25}) =$
 $2500 \pm 371.5 = (2128.5, 2871.5)$
- “I am 95% confident that this interval contains the true population mean.”

28

95% CI for Population Proportion

- A 95% confidence interval for the true population proportion is based on the sample proportion
- sample statistic $\pm Z_{0.05/2}$ (standard error)

$$\hat{p} \pm 1.96 \sqrt{\frac{\hat{p}\hat{q}}{n}}$$

29

Example: Student Ages > 40

- Take a sample of size $n = 4$
 - Sample: 28, 42, 43, 50
 - We can calculate the proportion of the sample that is greater than age 40 as $3/4 = 0.75$

30

Example: Student Ages > 40

- 95% confidence interval for the true population proportion
- 95% confidence interval (CI):
 $0.75 \pm 1.96 \sqrt{((0.75)(0.25)/4)} =$
 $0.75 \pm 0.42 = (0.33, 1.17)$
- “I am 95% confident that this interval contains the true population proportion.”

31

Sample Size and Estimation

- How is sample size related to estimation?
- If we are only interested in the precision of the estimate (e.g. a fairly narrow confidence interval) \Rightarrow
 - 1. Specify the width of the CI
 - 2. Make assumption about the standard deviation
 - 3. Solve for n

32

Sample Size for Single Mean

- Suppose we would like to estimate the true population mean age to within ± 5 years. Then
 - 1. Specify the width of the CI: ± 5
 - 2. Make assumption about the standard deviation:
assume $\sigma = 10$ based on previous data
 - 3. Solve for n

33

Sample Size for Single Mean

- Recall that
$$\bar{x} \pm z_{0.05/2} \frac{\sigma}{\sqrt{n}} = \bar{x} \pm 1.96 \frac{\sigma}{\sqrt{n}}$$
- Then we would set $5 = 1.96 \sigma/\sqrt{n}$
- Solving for n,
$$n = \frac{1.96^2 10^2}{5^2} = 15.4$$
- A sample size of 16 would allow us to estimate the true mean age to within ± 5 years

34

Sample Size for Single Proportion

- Suppose we would like to estimate the true population proportion over age 40 to within $\pm 3\%$. Then
 - Specify the width of the CI: ± 0.03
 - Make assumption about the true population proportion based on previous data (or use 0.5 as most conservative)
 - Solve for n

35

Sample Size for Single Proportion

- Recall that
$$\hat{p} \pm 1.96 \sqrt{\frac{\hat{p}\hat{q}}{n}}$$
- Then we would set $0.03 = 1.96 \sqrt{pq/n}$
- Using $p=0.5$ and solving for n,
$$n = \frac{1.96^2 (0.5)(0.5)}{0.03^2} = 1067$$
- A sample size of 1067 would allow us to estimate the true population proportion to within $\pm 3\%$

36

Summary

- Sampling distributions describe the theoretical distribution of the possible summary statistics obtained by sampling from a population
- Point estimates and confidence interval estimates quantify certainty
- Sample size estimates based on precision
