STA 291
Summer 2008

Lecture 8

Population Distribution

- Distribution from which we select the sample
- Unknown, we make inference about its parameters
- Mean =
- Standard Deviation =

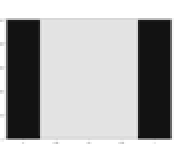
Sample Distribution

- Distribution of the data that we observe in the sample $X_1,\ldots,X_n$
- We use descriptive statistics to describe it
- If the sample size *n* increases, the sample distribution looks more and more like the population distribution
- Sample Mean =
- Sample Standard Deviation =

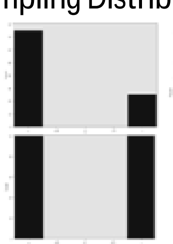
Sampling Distribution

- Probability distribution of a statistic (for example, the sample mean)
- Describes the pattern that would occur if we could repeatedly take random samples and calculate the statistic as often as we wanted
- Used to determine the probability that a statistic falls within a certain distance of the population parameter
- The mean of the sampling distribution of $\overline{X}$ is =
- The standard error of $\overline{X}$ is =

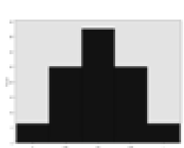
Summary: Population, Sample, and Sampling Distribution



Population distribution

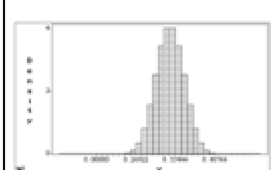


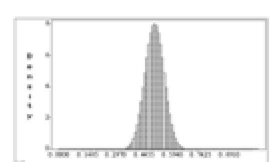
Two out of five possible sample distributions for samples of size *n*=4



Sampling Distribution for the sample mean for *n*=4

Sampling Distribution of the Sample Proportion for *n=25* and *n=100*





- Intuitively, larger samples yield more precise estimates
- *n=100* gives a narrower sampling distribution of the sample proportion than *n=25*
- Also, the histogram becomes more bell-shaped

Review: Population, Sample, and Sampling Distribution

Population distribution | One out of five possible sample distributions for samples of size *n*=4 | Sampling Distribution for the sample mean for *n*=4

•The population distribution is P(0)=0.5, P(1)=0.5.
•The sample distribution also takes the values 0 and 1, but the relative frequency depends on the sample chosen.
•The sampling distribution for the sample mean in a sample of size *n*=4 takes the values 0, 0.25, 0.5, 0.75, 1 with different probabilities.

Review: Population, Sample, and Sampling Distribution

- Population Distribution
  - Unknown distribution from which we select the sample
  - Want to make inference about its parameters
- Sample Distribution
  - Distribution of the data that we observe in the sample
  - We describe it, using descriptive statistics
  - For large *n*, it looks more and more like the population distribution
- Sampling Distribution
  - Probability distribution of a statistic (for example, the sample mean)
  - Used to determine the probability that a statistic falls within a certain distance of the population parameter
  - For large *n*, the sampling distribution of the sample mean looks more and more like a normal distribution

Central Limit Theorem

- The most important theorem in statistics
- For random sampling, as the sample size *n* grows, the sampling distribution of the sample mean $\bar{X}$ approaches a normal distribution
- This is the case even if the population distribution is discrete or highly skewed
- It is quite an amazing result

Central Limit Theorem

- Usually, the sampling distribution of $\bar{X}$ is approximately normal for *n*=20 or 30
- In addition, we know that the parameters of the sampling distribution are m and $s_{\bar{X}} = \frac{s}{\sqrt{n}}$
- For example:

If the sample size is *n*=25, then with 95% probability, the sample mean falls between

$$m - 1.96\frac{s}{\sqrt{n}} = m - \frac{1.96}{5}s \approx m - 0.4s$$

$$\text{and } m + 1.96\frac{s}{\sqrt{n}} = m + \frac{1.96}{5}s \approx m + 0.4s$$

$(m = \text{population mean}, s = \text{population standard deviation})$

Example

- Recall:
  - The scores on the Psychomotor Development Index (PDI) are approximately normally distributed with mean 100 and standard deviation 15. An infant is selected at random.
  - Find the probability that the infant's PDI score is at least 100.
  - Answer: 0.5
  - Find the probability that PDI is between 97 and 103.
  - Answer: 0.16
  - Find the *z*-score for a PDI value of 90. Would you be surprised to observe a value of 90?
  - Answer: -0.67; no, not surprised because 25% of the observations would even be below 90

Revised Example

- Refer to the previous exercise. A study uses a random sample of 225 infants
- Describe the sampling distribution of the sample mean PDI
- Find the probability that the sample mean falls between 97 and 103
- Find the *z*-score from the sampling distribution corresponding to a sample mean of 90 when the sample size is 225. Would you be surprised to observe a sample mean PDI of 90?
- Compare the results with those on the previous slide, and interpret the differences

Calculating *z*-Scores

1. **z-Score for an individual observation**
   – You need to know *X*, *mu*, and sigma to calculate *z*

$$z = \frac{X - m}{s}$$

2. **z-Score for a sample mean**
   – You need to know *X-bar*, *mu*, *sigma*, and *n* to calculate *z*

$$z = \frac{\bar{X} - m}{s / \sqrt{n}}$$

Chapter 10

- Statistical Inference: Estimation
  – Inferential statistical methods provide predictions about characteristics of a population, based on information in a sample from that population
  – For quantitative variables, we usually estimate the population mean (for example, mean household income)
  – For qualitative variables, we usually estimate population proportions (for example, proportion of people voting for candidate A)

Two Types of Estimators

- Point Estimate
  – A single number that is the best guess for the parameter
  – For example, the sample mean is usually at good guess for the population mean
- Interval Estimate
  – A range of numbers around the point estimate
  – To give an idea about the precision of the estimator
  – For example, "the proportion of people voting for A is between 67% and 73%"

Point Estimator

- A point estimator of a parameter is a sample statistic that predicts the value of that parameter
- A good estimator is
  – ***Unbiased***: Centered around the true parameter
  – ***Consistent***: Gets closer to the true parameter as the sample size gets larger
  – ***Efficient***: Has a standard error that is as small as possible

Unbiased

- An estimator is unbiased if its sampling distribution is centered around the true parameter
- For example, we know that the mean of the sampling distribution of "X-bar" equals "mu", which is the true population mean
- So, "X-bar" is an unbiased estimator of "mu"

Unbiased

- However, for any particular sample, the sample mean "X-bar" may be smaller or greater than the population mean
- "Unbiased" means that there is no systematic under- or overestimation
- If you repeatedly took samples, then the average of the sample means would converge to the population mean

Biased

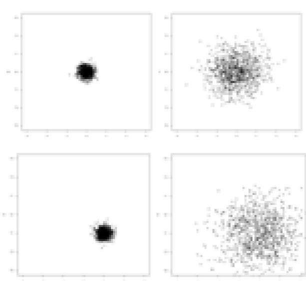
- A biased estimator systematically under- or overestimates the population parameter
- The definition of sample variance (and sample standard deviation) uses *n-1* instead of *n*, because this makes the sample variance unbiased
- With *n* in the denominator, it would systematically underestimate the variance

Efficiency

- An estimator is efficient if its standard error is small compared to other estimators
- Such an estimator has high precision
- A good estimator has ***small standard error and small bias*** (or no bias at all)

- The following pictures represent different estimators with different bias and efficiency
- Assume that the true population parameter is the point (0,0) in the middle of the picture

Bias and Efficiency



Note that even an unbiased and efficient estimator does not always hit exactly the population parameter.

But in the long run, it is the best estimator.

Point Estimators of the Mean and Standard Deviation

- The sample mean is unbiased, consistent, and (often) relatively efficient
- The sample standard deviation is unbiased when we use *n-1* in the denominator
- It is also consistent (and sometimes relatively efficient)

Example: Three Estimators

- Suppose we want to estimate the proportion of UK students voting for candidate A in the gubernatorial election
- We take a random sample of size $n=100$
- The sample is denoted $X_1, X_2, \ldots, X_n$, where $X_i=1$ if the *i*th student in the sample votes for A, $X_i=0$ otherwise

Example: Three Estimators

- Estimator 1 = the sample mean (sample proportion)
- Estimator 2 = the answer from the first student in the sample (X_1)
- Estimator 3 = 0.3
- Which estimator is unbiased?
- Which estimator is consistent?
- Which estimator has high precision (small standard error)?

Confidence Interval

- An inferential statement about a parameter should always provide the probable accuracy of the estimate
- How close is the estimate likely to fall to the true parameter value?
- Within 1 unit? 2 units? 10 units?
- This can be determined using the sampling distribution of the estimator/ sample statistic
- In particular, we need the standard error to make a statement about accuracy of the estimator

Confidence Interval

- If the sample size is n=25, then with 95% probability, the sample mean falls between

$$m - 1.96\frac{s}{\sqrt{n}} = m - \frac{1.96}{5}s \approx m - 0.4s$$

$$\text{and } m + 1.96\frac{s}{\sqrt{n}} = m + \frac{1.96}{5}s \approx m + 0.4s$$

$$(m = \text{population mean}, s = \text{population standard deviation})$$

Confidence Interval

- A confidence interval for a parameter is a range of numbers within which the true parameter likely falls
- The probability that the confidence interval contains the true parameter is called the confidence coefficient
- The confidence coefficient is a chosen number close to 1, usually 0.95 or 0.99

Confidence Interval

- The sampling distribution of the sample mean $\bar{X}$ has mean m and standard error

$$s_{\bar{X}} = \frac{s}{\sqrt{n}}$$

- If n is large enough, then the sampling distribution of $\bar{X}$ is approximately normal/bell-shaped

(Central Limit Theorem)

Confidence Interval

- To calculate the confidence interval, we use the Central Limit Theorem
- Therefore, we need sample sizes of at least, say, n=20
- Also, we need a z-score that is determined by the confidence coefficient
- Let's choose 0.95, then $z=1.96$

Confidence Interval

- With 95% probability, the sample mean falls in the interval between

$$m - 1.96\frac{s}{\sqrt{n}} \text{ and } m + 1.96\frac{s}{\sqrt{n}}$$

$$(m = \text{population mean}, s = \text{population standard deviation})$$

- Whenever the sample mean falls within 1.96 standard errors from the population mean, the following interval contains the population mean

$$\bar{X} - 1.96\frac{s}{\sqrt{n}} \text{ and } \bar{X} + 1.96\frac{s}{\sqrt{n}}$$

Confidence Interval

- So, the ***random*** interval between

$$\bar{X} - 1.96\frac{s}{\sqrt{n}} \text{ and } \bar{X} + 1.96\frac{s}{\sqrt{n}}$$

contains the population mean with 95% probability

- This is a confidence statement, and the interval is called a 95% confidence interval
- In practice, the population standard deviation is unknown and has to be replaced by its unbiased estimator, the sample standard deviation *s*

Confidence Interval

- A large sample 95% confidence interval for the population mean m is

$$\bar{X} \pm 1.96 \cdot \frac{s}{\sqrt{n}}$$

- where $\bar{X}$ is the sample mean and
- s is the sample standard deviation

Confidence Interval: Interpretation

- "Probability" means that "in the long run, 95% of these intervals would contain the parameter"
- If we repeatedly took random samples using the same method, then, in the long run, in 95% of the cases, the confidence interval will cover the true unknown parameter
- For one given sample, we do not know whether the confidence interval covers the true parameter
- The ***95% probability*** only refers to the ***method*** that we use, but not to the individual sample

Confidence Interval: Interpretation

- To avoid misleading use of the word "probability", we say:

"We are 95% confident that the true population mean is in this interval"

- Wrong statement:

"With 95% probability, the population mean is in the interval from 3.5 to 5.2"

Confidence Interval

- If we change the confidence coefficient from 0.95 to 0.99, the confidence interval changes
- Increasing the probability that the interval contains the true parameter requires increasing the length of the interval
- In order to achieve 100% probability to cover the true parameter, we would have to take the whole range of possible parameter values, but that would not be informative
- There is a tradeoff between precision and coverage probability
- *More coverage probability = less precision*

Confidence Interval

- Example: Find and interpret the 95% confidence interval for the population mean, if the sample mean is 70 and the sample standard deviation is 10, based on a sample of size

1. *n = 25*
2. *n = 100*