Notes 6: Sampling and Sampling Distributions

Julio Garín
Department of Economics

Statistics for Economics

Spring 2012

Introduction

- Recall from previous sections:
  - Element.
  - Population.
  - Sample.
- Motivation for sampling:
  - Bureau of Labor Statistics: unemployment rate surveys.
  - Proportion of voters supporting a candidate.
  - Outcome of a production process.
- We are interested in:
  - What constitutes a 'good' estimate?
  - What is the variance of the estimate?
  - What is the distribution of the estimate?

Sample

- Sample mean and sample proportion.
  - Examples.
- Sampling with and without replacement.

A *simple random sample* of size n from a finite population of size N is a sample selected such that each possible sample of size n has the same probability of being selected.

- What is the number of different simple random samples of size n that can be selected from a finite population of size N?

A *random sample* of size n from a infinite population is a sample selected such that the following conditions are satisfied:

1. Each element selected comes from the same population.
2. Each element is selected independently.

- Examples:
  - Quality control.
  - McDonald's customers or, more in general, customers arriving at a fast-food restaurants.

Point Estimation

- What is point estimation?
- Example: Sample of managers of a firm ($N = 2,500$).

Table: Annual Salary and Training Program Status for a Simple Random Sample of 30 Managers

Salary ($)	MGMT Training Program	Salary ($)	MTP
$x_1 = 49,094.30$	Yes	$x_{16} = 51,766.00$	Yes
$x_2 = 53,263.90$	Yes	$x_{17} = 52,541.30$	No
.	.	.	.
.	.	.	.
$x_{15} = 53,188.20$	No	$x_{30} = 57,309.10$	No

- From this sample we can estimate $\bar{X}$, s^2, and $\bar{p}$

Point Estimation

- From our population and our sample we obtain:

Table: Summary of point estimates and population parameters

Population Parameters	Value	Point Estimator	Estimate
μ	\$51,800	$\bar{x}$	\$51,814
σ	\$4,000	s	\$3,348
p	0.60	$\bar{p}$	0.63

- Why population parameters and point estimator differ?
- Importance of having a close correspondence between the sampled population and the *target population*.
  - Sometimes they are not the same: Amusement Park with restricted attendance.
- Why is this important?

Sampling Distributions

Introduction to Sampling Distributions

- In a survey:
  - n people are surveyed.
  - $x_1, x_2, \cdots, x_n$ are the responses.
  - Underlying population has expected value μ and variance σ^2.
- Suppose we are interested in $\bar{x}$
- $\bar{x}$ will be a random variable itself.
  - Subject to sampling error.
  - Will be different each time, based on who is in the sample.
- Since $\bar{x}$ is a r.v. it will have a mean, a standard deviation, and a probability distribution.

Sampling Distribution: Example

Table: Values of $\bar{x}$ and $\bar{p}$ from 500 Random Samples of 30 Managers

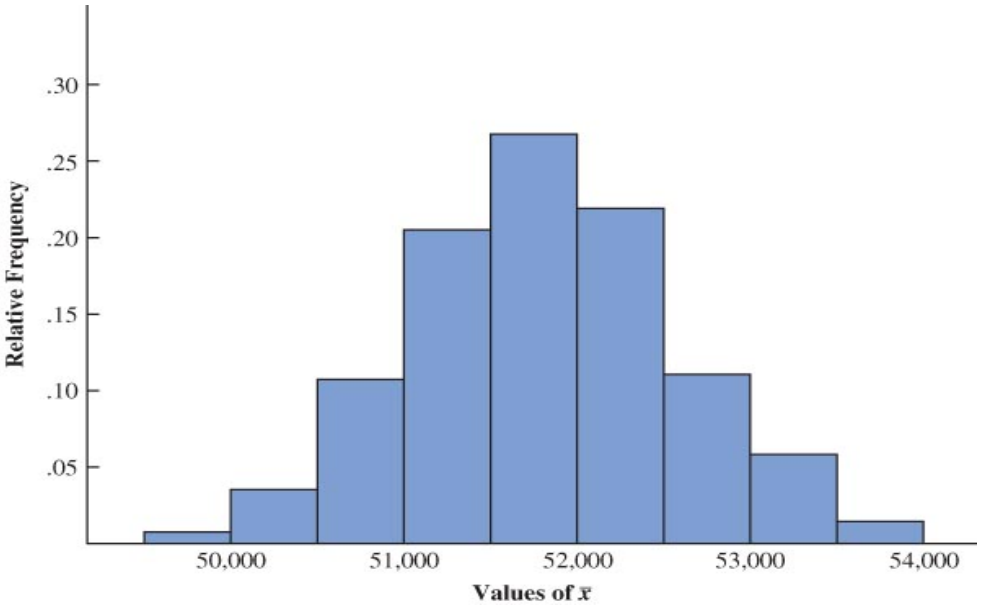
Sample Number	Sample Mean	Sample Proportion
1	\$51,814	0.63
2	\$52,670	0.70
3	\$51,780	0.67
.	.	.
.	.	.
.	.	.
500	\$51,752	0.50

The probability distribution of a point estimator is called the *sampling distribution* of that estimator.

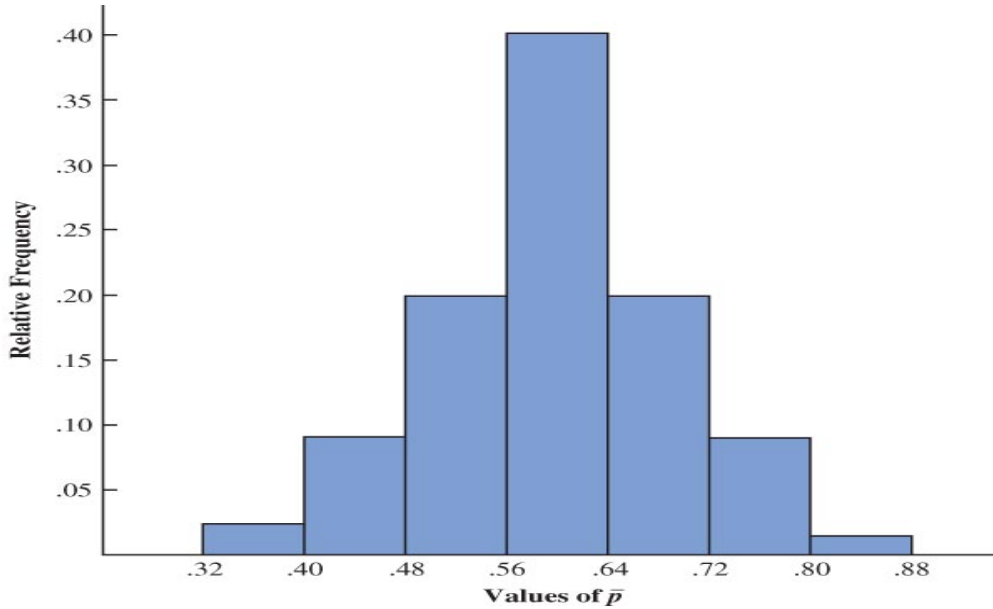
- Why is the sampling distribution important?

Relative Frequency Histogram From 500 Simple Random Samples ($n = 30$)

Histogram of $\bar{x}$ values.



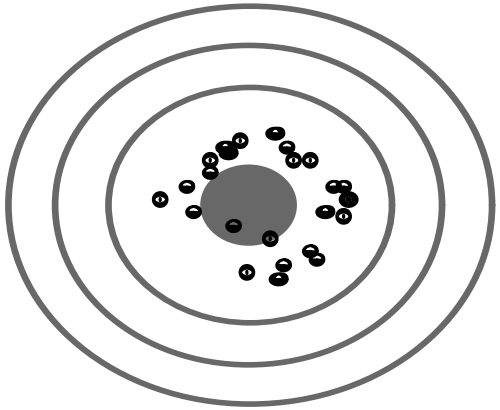
Histogram of $\bar{p}$ values.



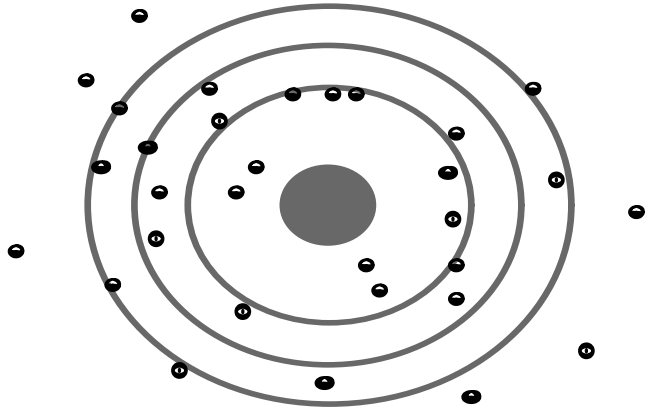
Selecting Good Estimators

- Given our goal, what constitutes a 'good' estimate of a population parameter θ?
- We will use one criteria:
  - Unbiased estimates.
    - Mathematically?
- When is an estimate unbiased?
- Important: unbiased does not mean that you get the "correct" answer every time!
  - Bias and precision.

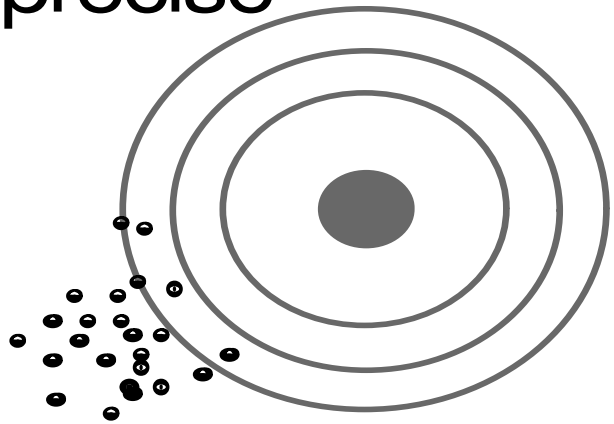
Difference Between Precision and Bias



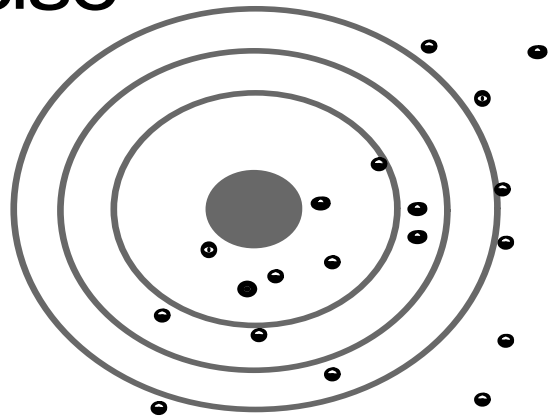
Unbiased and precise



Unbiased but not precise



Biased but precise



Biased and not precise

Example of Biased Estimates: Predicting Election Results

- Want to survey people about and upcoming election.
  - Not all people can vote.
  - Not all people who can vote will.
- If you want to accurately predict the outcome, you must include "likely" voters.
- Suppose there are two 'facts':
  - Democrats prefer democratic candidates.
  - Democrats are less likely to vote.
- Problems?
- How do surveys handle this problem?
  - Survey registered voters.
  - Collect demographic data.
  - Use data from previous elections.

Sampling Distribution of $\bar{x}$

Expected Value of $\bar{x}$

- Why we need an expected value of a mean?
- Recall:

The sampling distribution of $\bar{x}$ is the probability distribution of all possible values of the sample mean $\bar{x}$.

The expected value of $\bar{x}$, $\mathbb{E}(\bar{x})$ is

$$\mathbb{E}(\bar{x}) = \mu$$

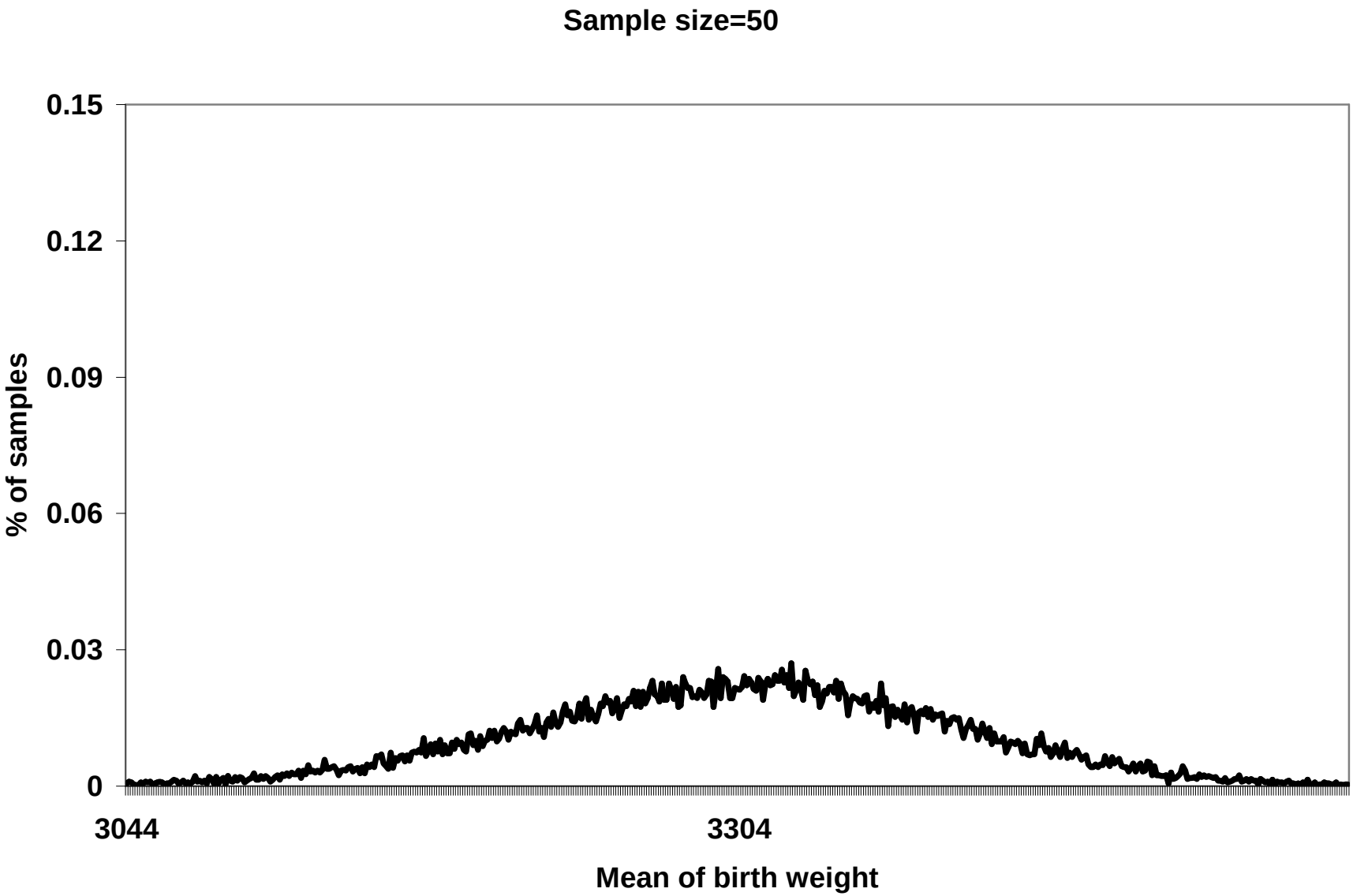
where μ is the population mean.

- First question: is $\bar{x}$ an unbiased point estimator?

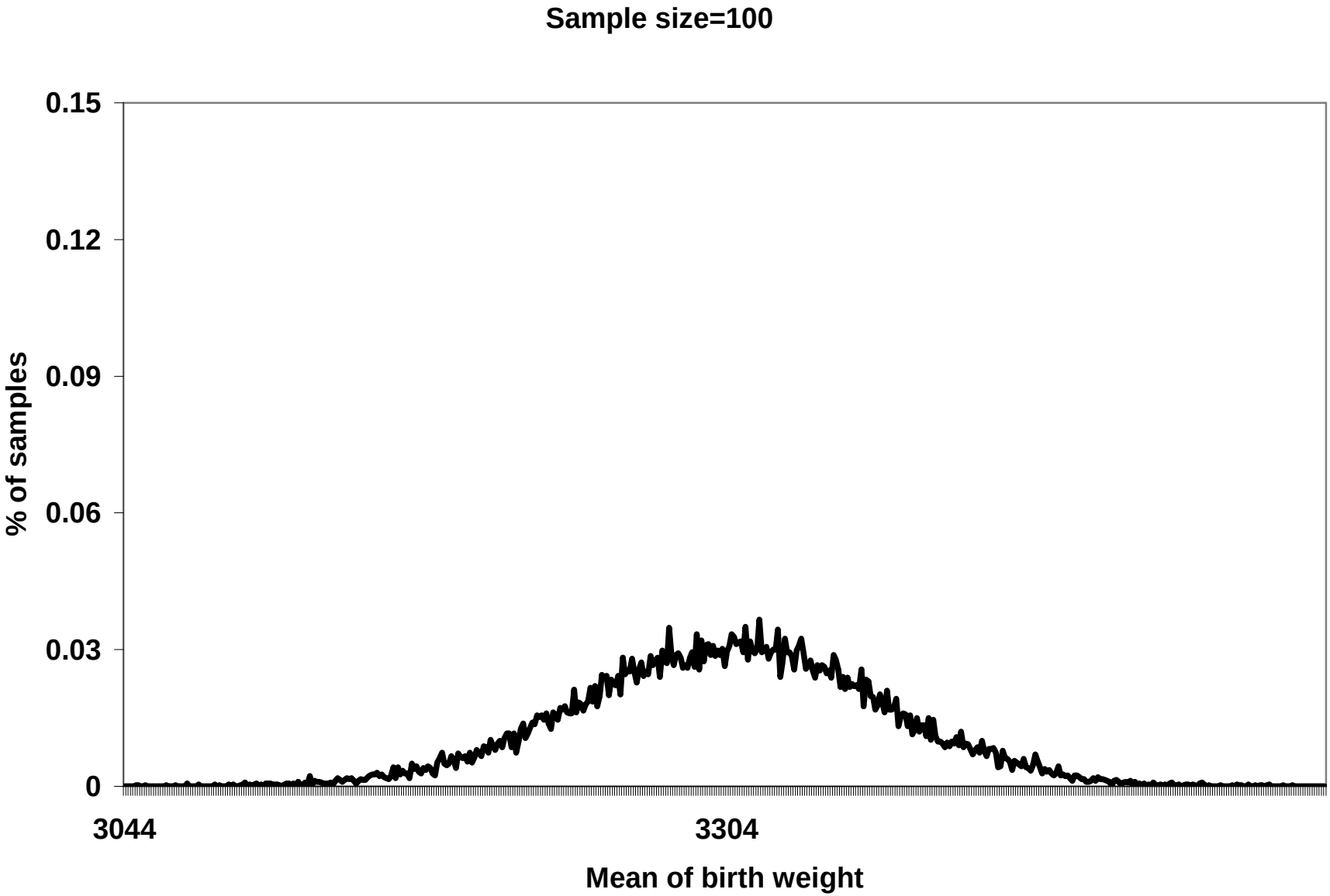
Example: Births in the State of Maryland

- Consider a population: all births in the state of Maryland in 1997 ($N = 65,990$).
- The variable of interest is the birth weight of children. From the population:
  - $\mu = 3,304$ grams.
  - $\sigma = 651$ grams.
- Consider now two different ways of constructing a sampled mean birth weight.
  - Sample 5 births.
  - Sample 50,000 births.
- What would you expect to see in their respective sampling distributions?
- Our exercise:
  - Construct mean with samples of 50, 100, 500, 1,000, 2,000 randomly selected people.
  - Construct 25,000 of these means.
  - Graphs these distributions.

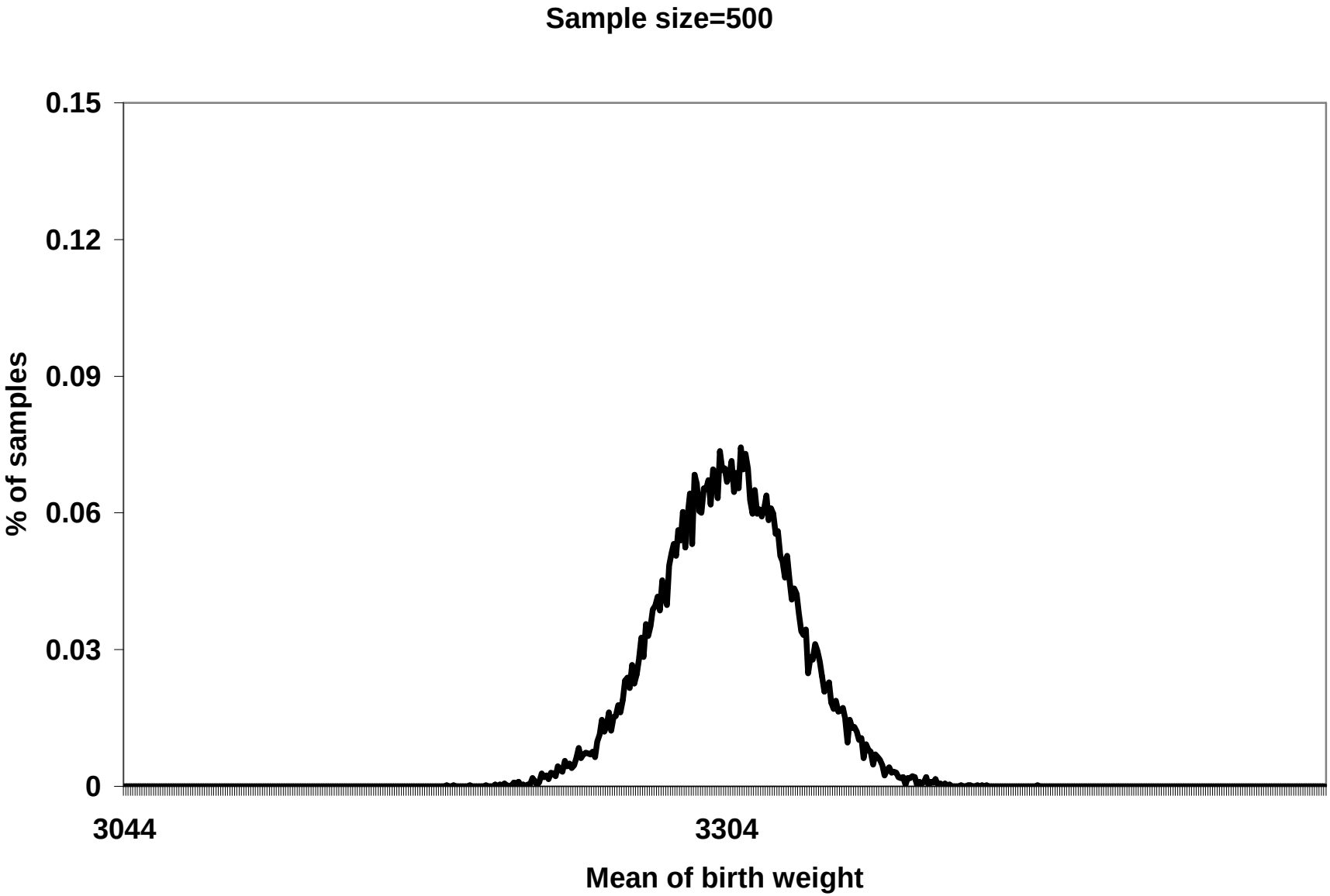
Birth Weight of Children in Marylan ($n = 50$)



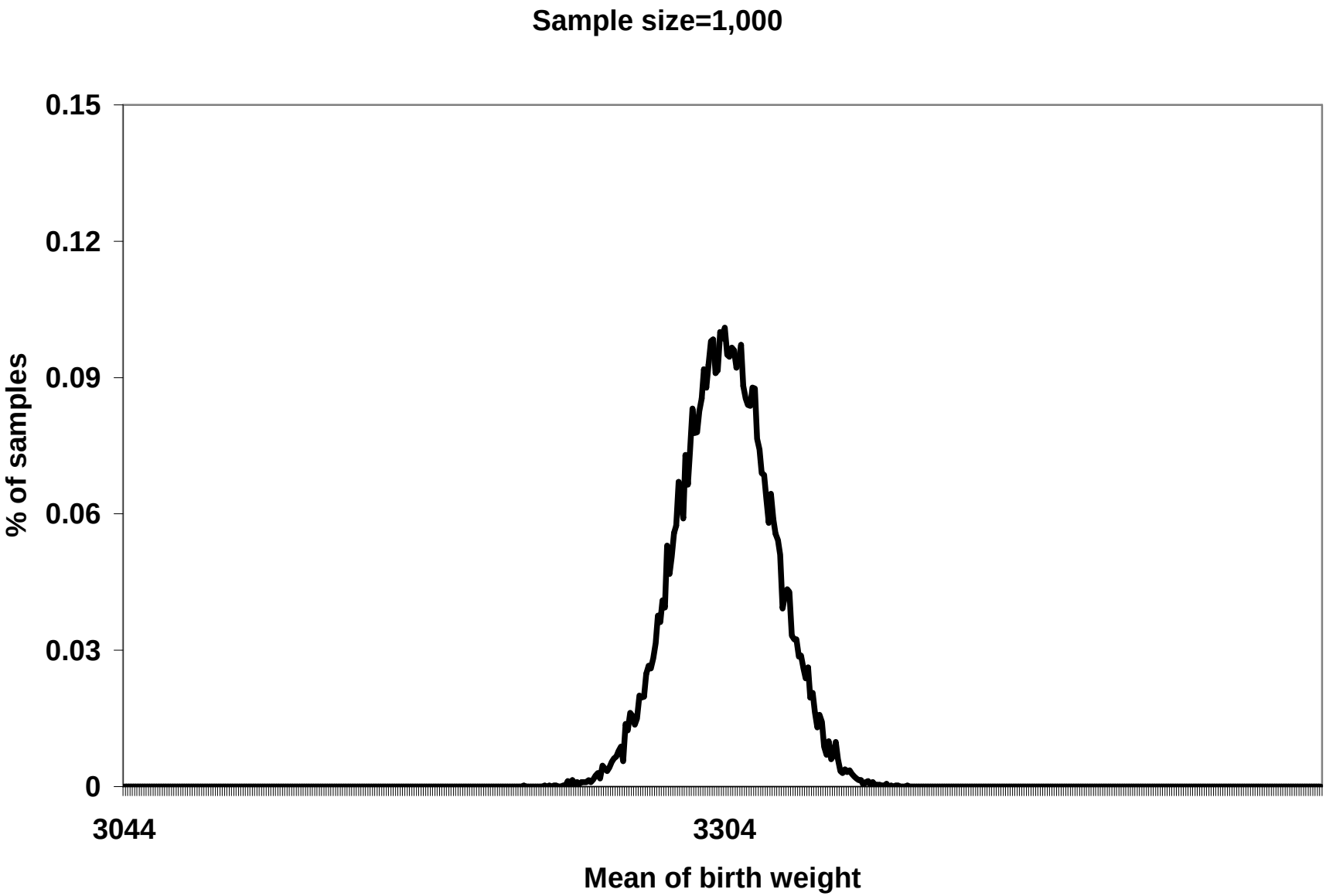
Birth Weight of Children in Maryland ($n = 100$)



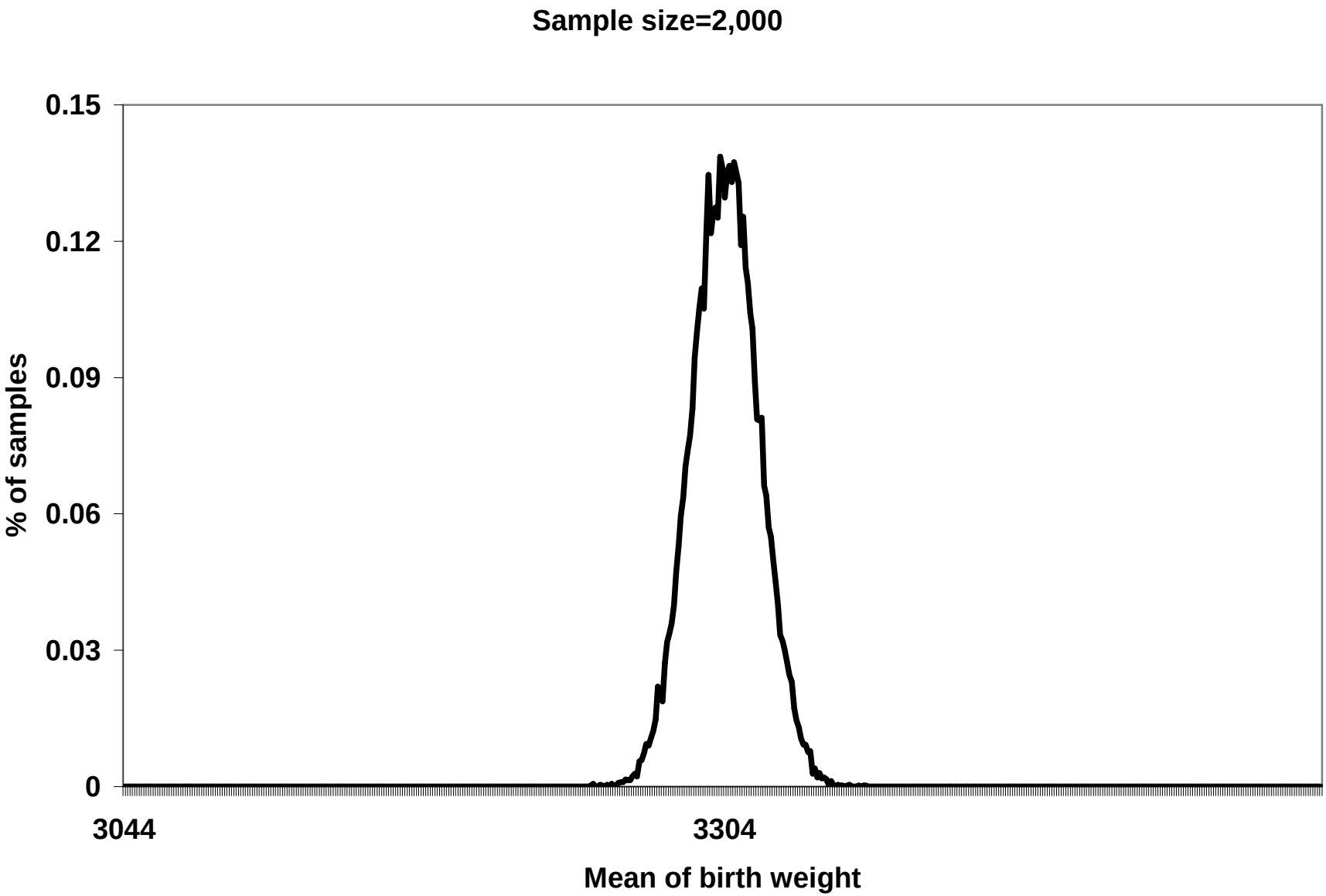
Birth Weight of Children in Maryland ($n = 500$)



Birth Weight of Children in Maryland ($n = 1,000$)



Birth Weight of Children in Maryland ($n = 2,000$)



Birth Weight: Summary of the Plots

Table: Distribution of Mean Birth Weight

Size of Random Sample (n)	**Mean of $\bar{x}$**
50	3,303.3
100	3,304.9
500	3,304.1
1,000	3,304.1
2,000	3,304.3

Variance of $\bar{x}$

- What would be the variance of $\bar{x}$?
- n **independent** draws: $x_1, x_2, \cdots, x_n$.
- We could use the Mean Squared Deviation (MSD):

$$MSD = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

- Is this an unbiased estimator of the population variance?
- Let us derive the variance:
  - $\bar{x}$ is a linear combination of independent random variables...
  - For random variables X and Y, what is the variance of $Z = \alpha + \beta_1 X + \beta_2 Y$?
  - Using independence.
  - Let us extend this definition to combination of n independent random variables.

Variance of $\bar{x}$

- Since the MSD is biased, what would be an unbiased estimator of the variance?
  - Our friend: the sample variance, s^2!
  - You will have to prove is unbiased in a homework question.

The standard deviation of $\bar{x}$ is

$$\sigma_{\bar{x}} = \left(\frac{\sigma}{\sqrt{n}} \right)$$

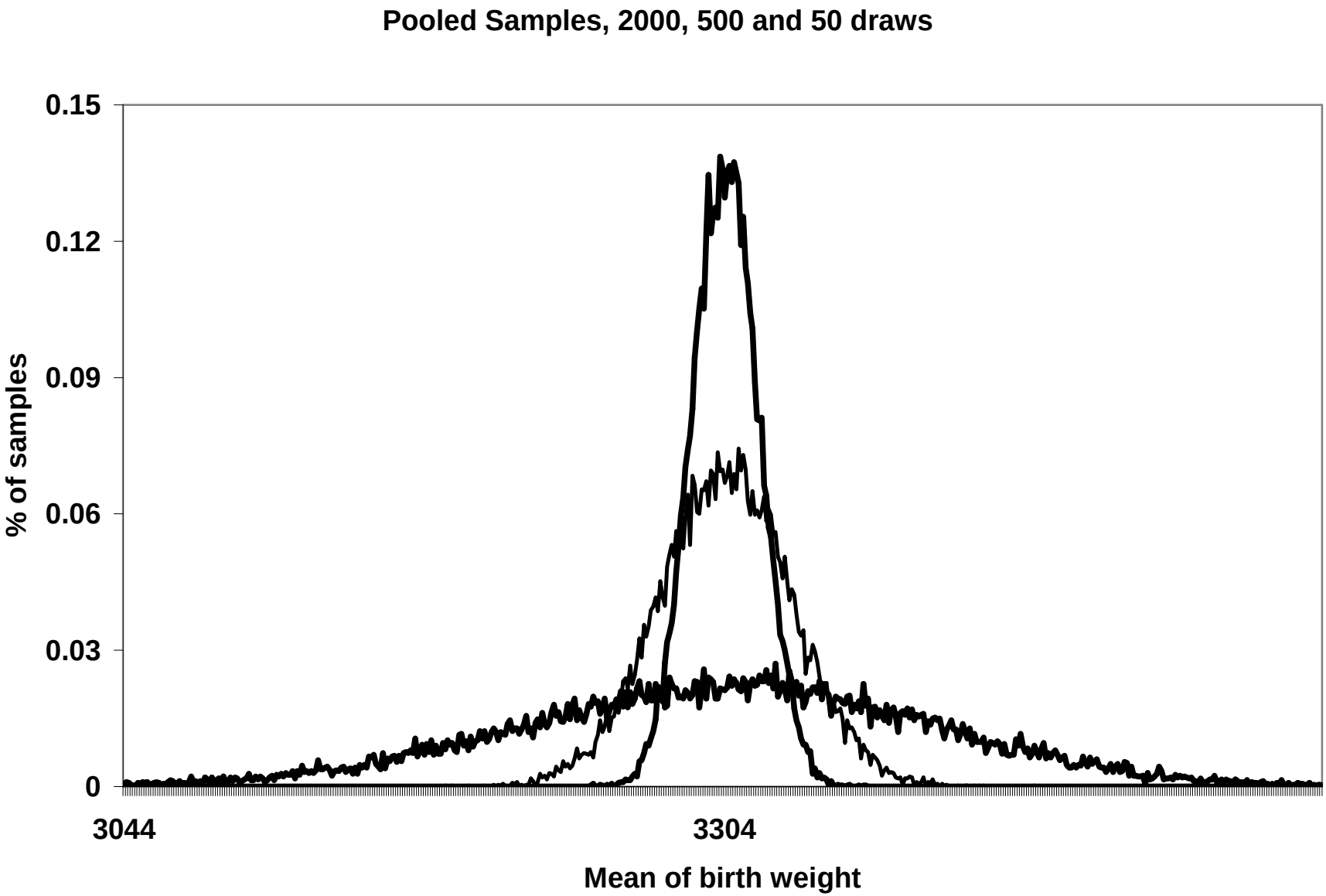
- We will refer to $\sigma_{\bar{x}}$ as the *standard error* of the mean.
- What will be the effect of increasing the sample?

Relationship Between Sample Size and The Sampling Distribution of $\bar{x}$

Table: Distribution of Mean Birth Weight

Size of R.S. (n)	**Mean of $\bar{x}$**	$\frac{\sigma^2}{n}$	$\sigma_{\bar{x}}$
50	3,303.3	8,476	8,381
100	3,304.9	4,238	4,239
500	3,304.1	848	828
1,000	3,304.1	424	417
2,000	3,304.3	212	209

Variance and Sample Size, Graphically



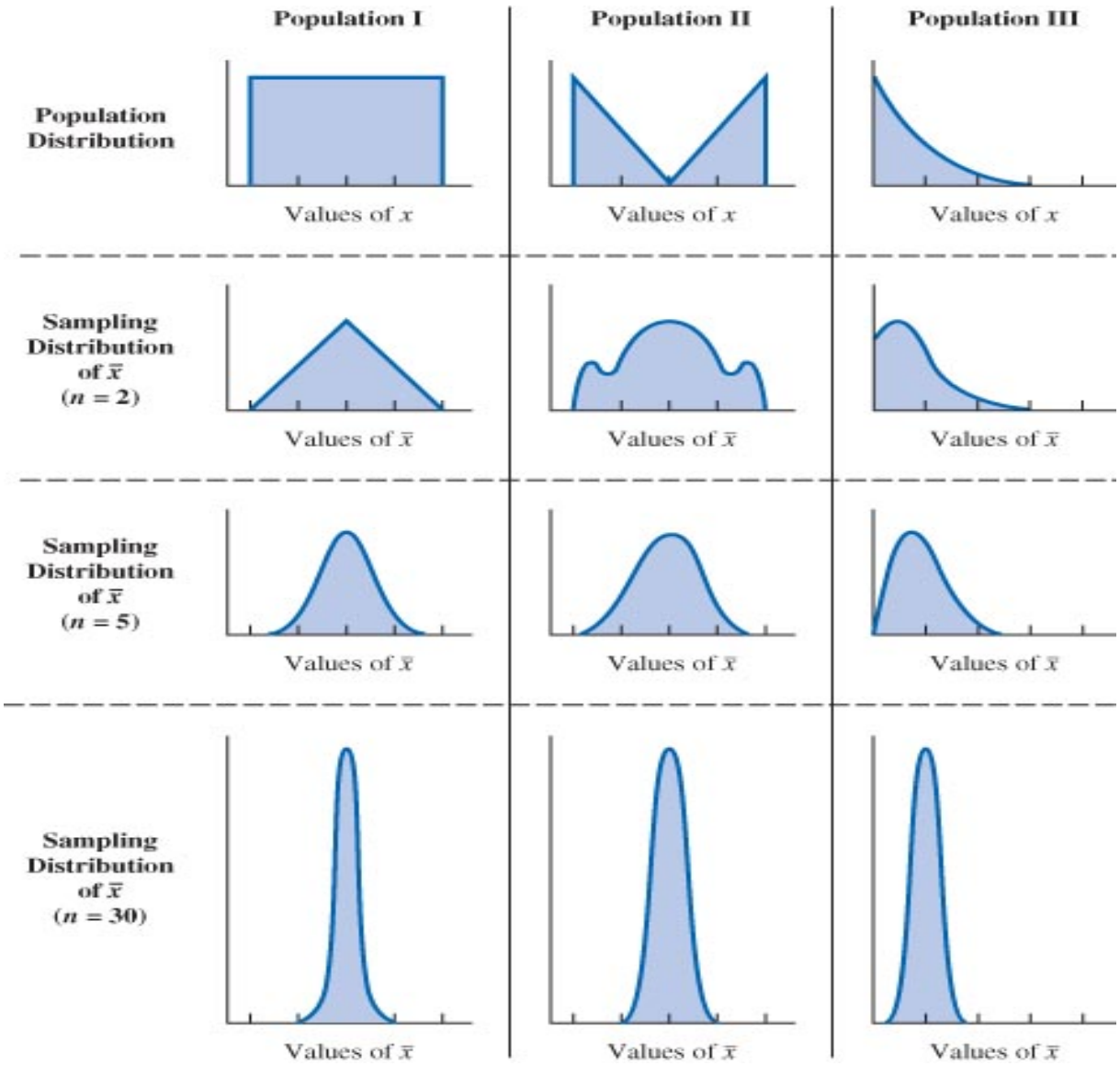
Shape of the Sampling Distribution of $\bar{x}$

- Now that we have the first two moments. What is the form of the sampling distribution?
- Two possibilities:
  1. The population has a normal distribution.
  2. The population does not have a normal distribution.

Central Limit Theorem (CLT): Given n draws to a population with mean μ and variance σ^2, the distribution of $\bar{x}$ approaches a normal distribution with mean μ and variance σ^2/n

$$\bar{x} \sim \mathcal{N}(\mu, \sigma^2/n) \quad \text{as } n \text{ increases}$$

Illustration of the CLT



How Can We Utilize the CLT?

- We could answer questions about the reliability of your survey.
  - What is the probability that $\bar{x}$ will fall in a specific range?
  - How large does the sample size have to be to maintain a certain level of reliability?

Application: Prepare for the Worst

A particular elevator will collapse if the interval weight is in excess of 2,080 pounds. The population values for weight in the population are $\mu = 150$ and $\sigma = 25$. The manufacturer wants to restrict the number of people on the elevator at one time to prevent an accident.
Suppose they restrict the occupancy to a maximum of 13 people.

1. What is the chance that if 13 people enter the elevator, the weight will exceed 2,080 pounds?
2. What if we decrease the occupancy limit to 11 people?
3. Conclusions?

Example: Golfers

The average score for male golfers is 95 and the average score for female golfers is 106 (*Golf Digest*, April 2006). Use these values as the population means for men and women and assume that the population standard deviation is $\sigma = 14$ strokes for both. A simple random sample of 30 males golfers and another simple random sample of 45 females golfers will be taken.

1. Show the sampling distribution of $\bar{x}$ for male golfers.
2. What is the probability that the sample mean is within three strokes of the population mean for the sample of male golfers?
3. What is the probability that the sample mean is within three strokes of the population mean for the sample of female golfers?
4. In which case, part b) or part c), is the probability of obtaining a sample mean within three strokes of the population higher? Why?

Accuracy and Sample Size

Suppose in a population of high school seniors, SAT Math scores have an expected value of 550 and a standard deviation of 100 (population values). Suppose you survey 25 students and get a sample mean of $\bar{x} = 540$.

- What is the standard error?
- Why it is so high?
- What we need to do in order to cut the standard error in half?

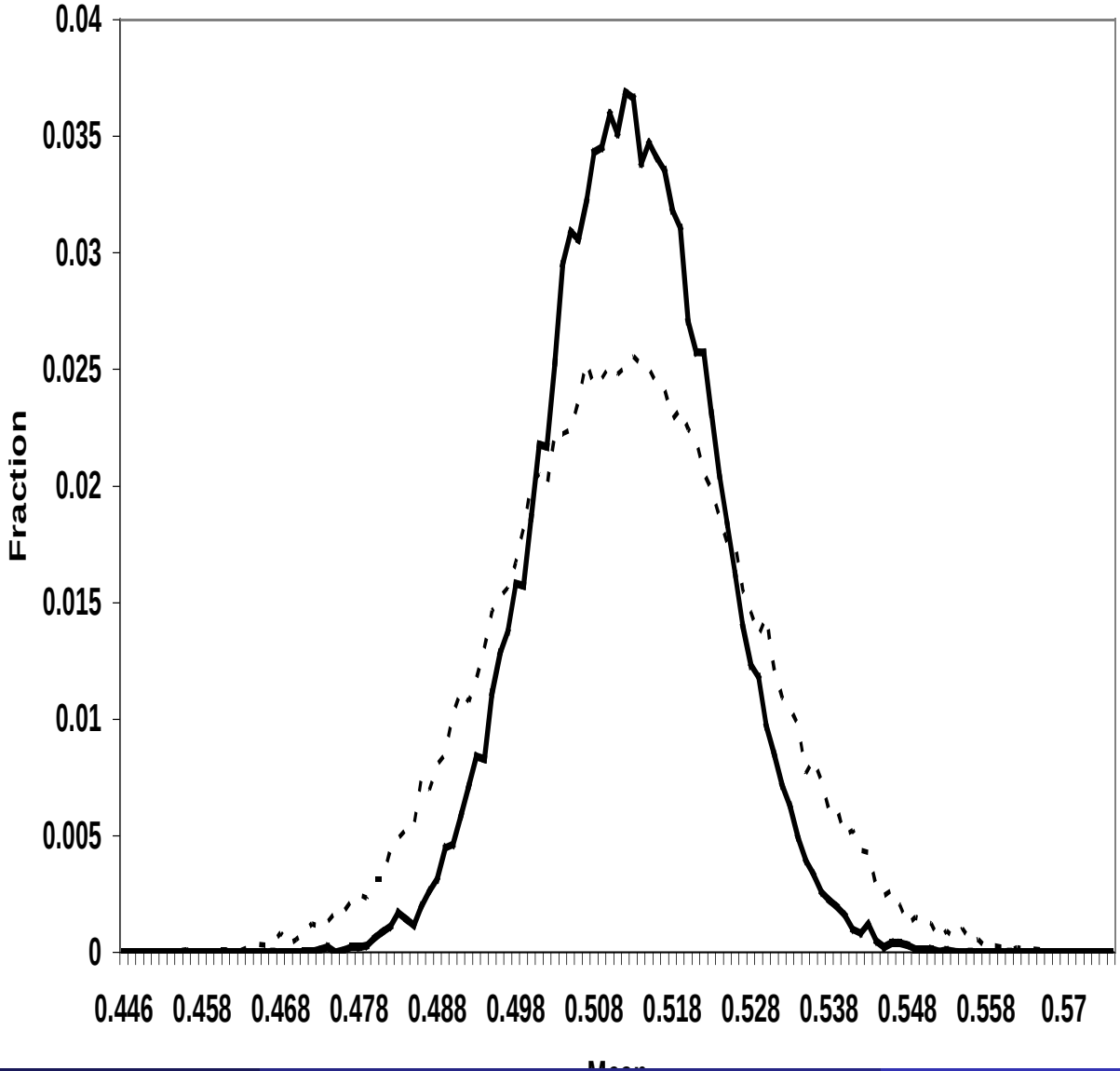
In general, to double accuracy (as measured by standard error) we need to quadruple the sample size.

Discrete Outcome

Discrete Outcome

- What if the outcome is discrete?
  - The definition of standard error remains the same.
- Outcome: Whether the birth is a boy.
  - Equal to 1 if boy, zero otherwise.
  - Assume the probability of success is 0.512.
  - How this is distributed?
  - Expected value?
  - Variance of this fraction?
  - What we do expect to be the variance of fraction male with 500 draws?

Distribution of Sample Mean, Fraction of Boys (samples of 2,000 and 1,000)



Proportions

- $\text{Var}(\bar{x})$ for a binomial distribution?
- Remember, we are referring to proportions here.
- What is the variance of the sample proportion?

The sampling distribution $\bar{p}$ is the probability of all possible values of the sample proportion $\bar{p}$. Where $\bar{p} = \frac{x}{n}$ where x is the number of elements in the sample that possess the characteristic of interest.

- What is the practical value of having this sampling distribution?

The expected value of $\bar{p}$ is $\mathbb{E}(\bar{p}) = p$

- Remember $np \geq 5$ and $n(1-p) \geq 5$ rule?

Example: Variance of the Proportion

A group of doctors is investigating the medical conditions of newborns. They collect data from the next 30 babies born. The fraction of girls in the sample is 47%.

1. What is the standard error of this estimate?

Example I: Better Business Bureau

In 2008 the Better Business Bureau settled 75% of complaints it received (*USA Today*, March 2, 2009). Suppose you have been hired by the Better Business Bureau to investigate the complaints it received this year involving new car dealers. You plan to select a sample of new car dealer complaints to estimate the proportion of complaints the BBB is able to settle. Assume the population proportion of complaints settled for new car dealers is 0.75, the same as the overall proportion complaints settled in 2008.

1. Suppose you select a sample of 450 complaints involving new car dealers. Show the sampling distribution of $\bar{p}$.
2. Based upon a sample of 450 complaints, what is the probability that the sample proportion will be within 0.03 of the population proportion?
3. Suppose you select a sample of 200 complaints involving new car dealers. Show the sampling distribution of $\bar{p}$.
4. Based upon the smaller sample of only 200 complaints, what is the probability that the sample proportion will be within 0.03 of the population proportion?

Exercise II: Sample Reliability

Weekly earnings have the following moments:

$$\mu = 737\$, \quad \sigma = 701\$$$

Suppose you survey 100 people.

1. What is the chance $\bar{x}$ will be within \$50 of the 'true' value, μ?
2. Is this survey accurate?

Exercise III: Smoking Rates

Most national surveys estimate the adult smoking rate to be between 24 and 28 percent, with a population expected value of 26 percent. Suppose you survey 500 people.

1. What is the chance that $\bar{p}$ will lie in this range?

Exercise IV: Sales Representative

J. is an awesome, good looking, and humble Uruguayan soccer player and an even better sales representative for a major publisher of college textbooks. Historically, Mr. J. obtains a book adoption of 25% of his sales calls. Viewing his sales calls for one month as a sample of all possible sales calls, assume that a statistical analysis of the data yields a standard error of the proportion of 0.0625.

1. How large was the sample used in the analysis? That is, how many sales calls did Mr. J. make during the month?
2. Let $\bar{p}$ indicate the sample proportion of book adoptions obtained during the month. Show the sampling distribution of $\bar{p}$.
3. Using the sampling distribution of $\bar{p}$, compute the probability that J. will obtain book adoptions on 30% or more of his sales during a one-month period.

Exercise V: Picking Optimal Sample Sizes

Suppose we know that family income in 1997 averaged \$32K with a standard deviation of \$17K. You would like to survey families about their income, but you do not want mean family income to vary too much from sample to sample.

1. How many families do you have to survey so that $\bar{x}$ is within \$4K the population mean 95% percent of the time?
2. What if we want $\bar{x}$ to be within \$1,000 95% of the time?

Exercise VI: Two Shorts

A market research firm conducts telephone surveys with a 40% historical response rate.

1. What is the probability that in a new sample of 400 telephone numbers, at least 150 individuals will cooperate and respond to the questions?

The mean television viewing time for American is 15 hours per week (*Money*, November 2003). Suppose a sample of 60 Americans is taken to further investigate viewing habits. Assume the population standard deviation for weekly viewing time is $\sigma = 4$ hours.

1. What is the probability that the sample mean will be within 1 hour of the population mean?
2. What is the probability that the sample mean will be within 45 minutes of the population mean?

Material Covered So Far

Review.