

# Statistics for IT Managers

## 95-796, Fall 2007

### Course Overview

Instructor: Daniel B. Neill ([neill@cs.cmu.edu](mailto:neill@cs.cmu.edu))

TAs: Chirantan Chaterjee ([chirantan@cmu.edu](mailto:chirantan@cmu.edu))  
Chris Harle ([charle@andrew.cmu.edu](mailto:charle@andrew.cmu.edu))

# Statistics: why bother?

We have some problem we want to solve:

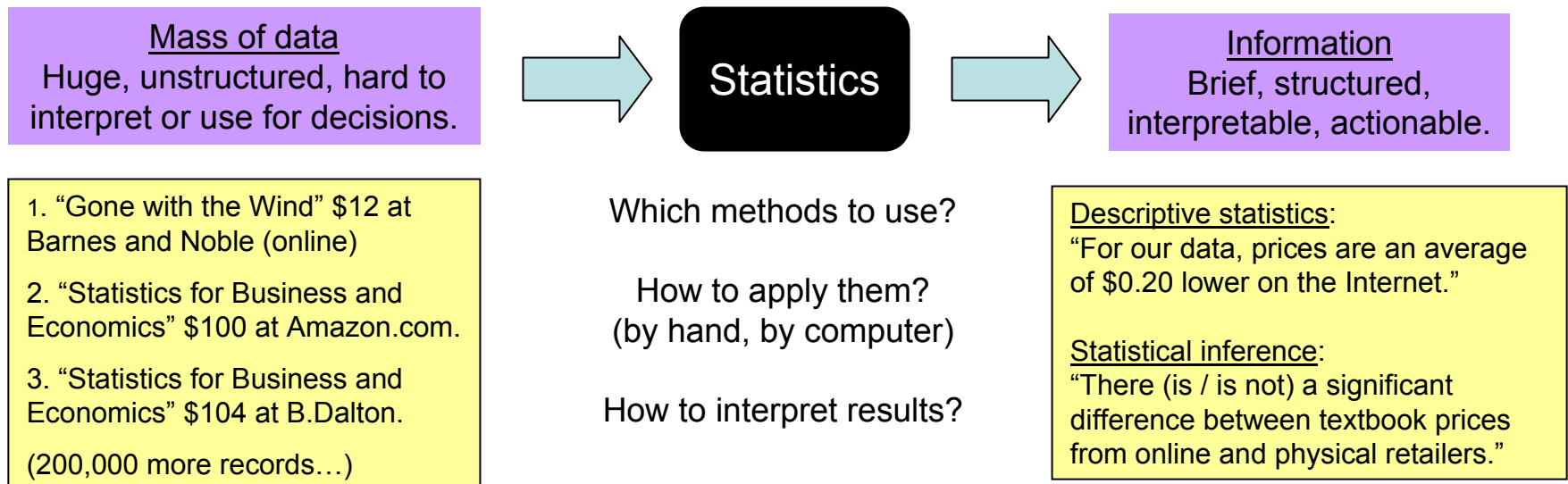
“Are book prices lower on the Internet?”

“What industry sectors are most profitable?”

“Should we invest in a new technology?”

Option 1: Rely on intuition (“Because users can more easily compare prices on the Internet, this will lead to more price competition and thus lower prices.”)

Option 2: Collect and analyze real-world data to test whether your intuitions are correct.



# Goals of the course

- To provide individuals who aspire to IT management positions with the basic statistical tools for analyzing and interpreting data.
- By the end of this course, you should be able to correctly choose and apply the appropriate statistical methods for real-world problems related to IT management.
- Because most real-world datasets are too large to analyze by hand, you will be expected to learn and use the statistical software package Minitab.

# Structure of the course

- 14 lectures divided into three modules:
  - Descriptive statistics and probability (4 lectures)
  - Hypothesis testing and inference (6 lectures)
  - Simple and multiple regression (4 lectures)
- Grades will be based on:
  - Three homeworks 30% (10% each)
  - Two mini-projects 30% (15% each)
  - Final exam 40%
- See syllabus on Blackboard for detailed schedule, and for course policies (cheating, late work, re-grades).

# Course textbook and slides

- Statistics for Business and Economics (10<sup>th</sup> ed.) by McClave, Benson, and Sincich.
  - Module 1 (Descriptive statistics and probability) covers Chapters 1-4.
  - Module 2 (Statistical inference) covers Chapters 5-7.
  - Module 3 (Regression) covers Chapters 10-11.
- Not all sections of these chapters will be covered. See syllabus for readings corresponding to each lecture.
- Slides for each module will be available ahead of time on Blackboard.

# Statistics for IT Managers 95-796, Fall 2007

## Module 1: Descriptive Statistics and Probability (4 lectures)

Reading: Statistics for Business and Economics, Ch. 1-4

\*\* Mini-project 1 due Monday 9/10 \*\*

\*\* Homework for Module 1 due Monday 9/17 \*\*

# Basic definitions

- **Statistics** is the science of analyzing and interpreting data, i.e. transforming raw data into information.
- **Descriptive statistics** are used to organize and summarize data, and to present this information in a convenient and usable form.
  - Graphical displays (e.g. histograms, box plots)
  - Numerical summaries (e.g. mean, median, mode, variance)
- **Inferential statistics** use sample data to make estimates, decisions, predictions, or other generalizations about a larger set of data.
  - **Population:** data measuring some characteristic of all members of a group (“all teenage males who watch television”)
  - **Sample:** data on a representative subset of the population (“100 randomly sampled teenage males who watch television”)

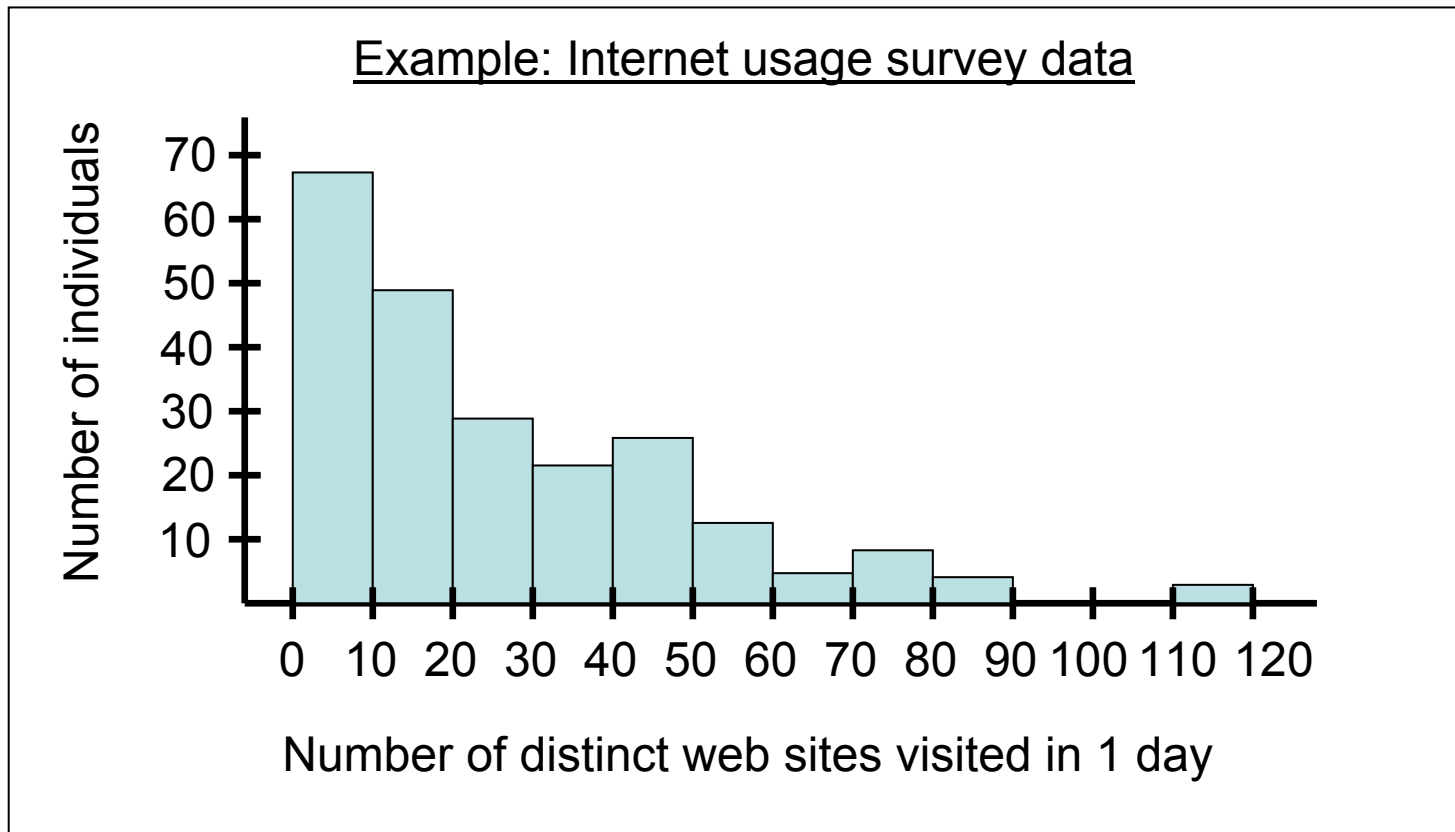
What can we conclude about the population, based on our sample?

# Data types

- **Qualitative (or categorical) data:** each data point is classified into one of a given set of categories.
  - **Nominal data:** categories do not have a given order.
    - Animal type: {dog, cat, bird, fish}.
  - **Ordinal data:** categories have a given order.
    - Movie ranking: 1-5 stars.
- **Quantitative (or numerical) data:** each data point is measured on a naturally occurring numerical scale.
  - Height, weight, income, etc.

# Histograms

- One of the many graphical methods for displaying numerical data.
- Shows counts or percentages of data in each interval.



# Numerical descriptive statistics

- **Measures of the center of the data**
  - Mean, median, mode
- **Measures of variability**
  - Variance, standard deviation, range, interquartile range
- **Some advantages of numerical statistics:**
  - More succinct than graphical methods
  - Less subject to distortion
  - Form the basis for statistical inferences
- Any disadvantages?

# Measures of the center

- **Mean:** the average of all values.

$$\bar{x} = \frac{\sum x_i}{n} = \frac{(x_1 + x_2 + \dots + x_n)}{n}$$

$x_i$  = value of the  $i^{\text{th}}$  observation  
 $n$  = total number of observations

- **Median:** the “middle” number when measurements are arranged in ascending (or descending) order.
- **Mode:** the most common value.

Example dataset: 1, 1, 2, 2, 2, 3, 4, 4, 5, 16

$$\text{Mean} = (1 + 1 + 2 + 2 + 2 + 3 + 4 + 4 + 5 + 16) / 10 = 4$$

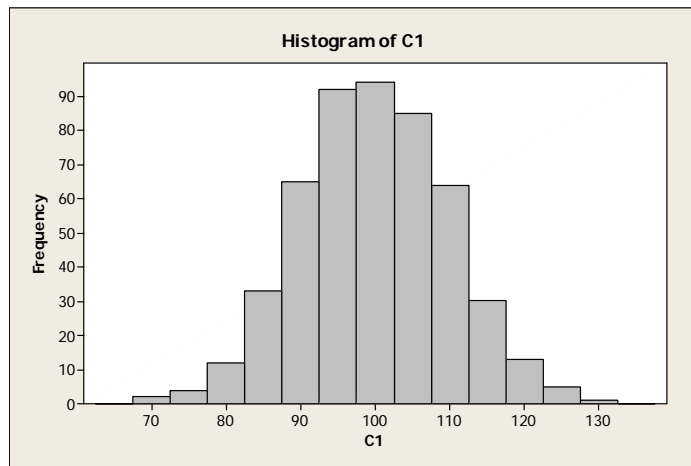
$$\text{Median} = (2 + 3) / 2 = 2.5$$

$$\text{Mode} = 2$$

Notice that the mean is more affected by outlier values than the median!

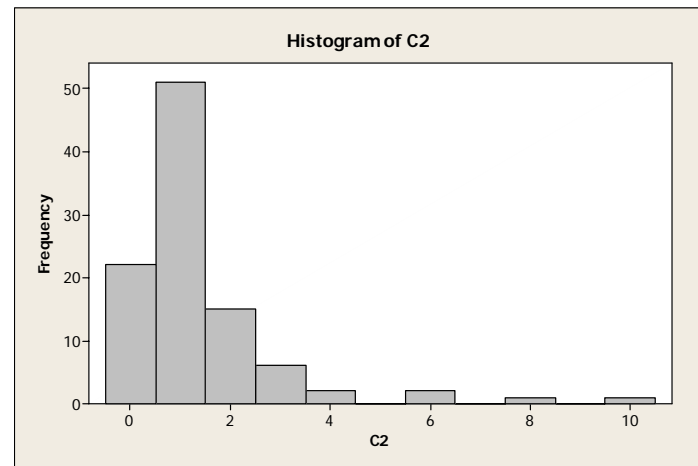
# Skewed distributions

- A distribution is **symmetric** if mean = median.
- A distribution is **positively skewed** if mean > median.
- A distribution is **negatively skewed** if mean < median.



500 values generated from  $N(100, 10)$   
Mean = 99.83, Median = 99.91

Approximately symmetric

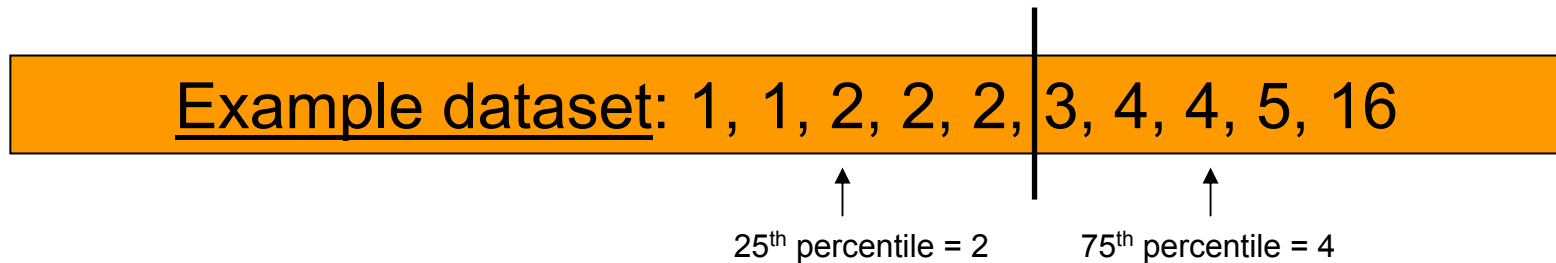


100 values generated from  $F(3, 5)$   
Mean = 1.37, Median = 0.88

Positively skewed

# Measures of variability

- **Range:** the difference between the smallest and largest observations.
- **Interquartile range:** the difference between the 25<sup>th</sup> and 75<sup>th</sup> percentiles, where the k<sup>th</sup> percentile is a value such that k% of the observations are below that value and (100-k)% of the observations are above that value.



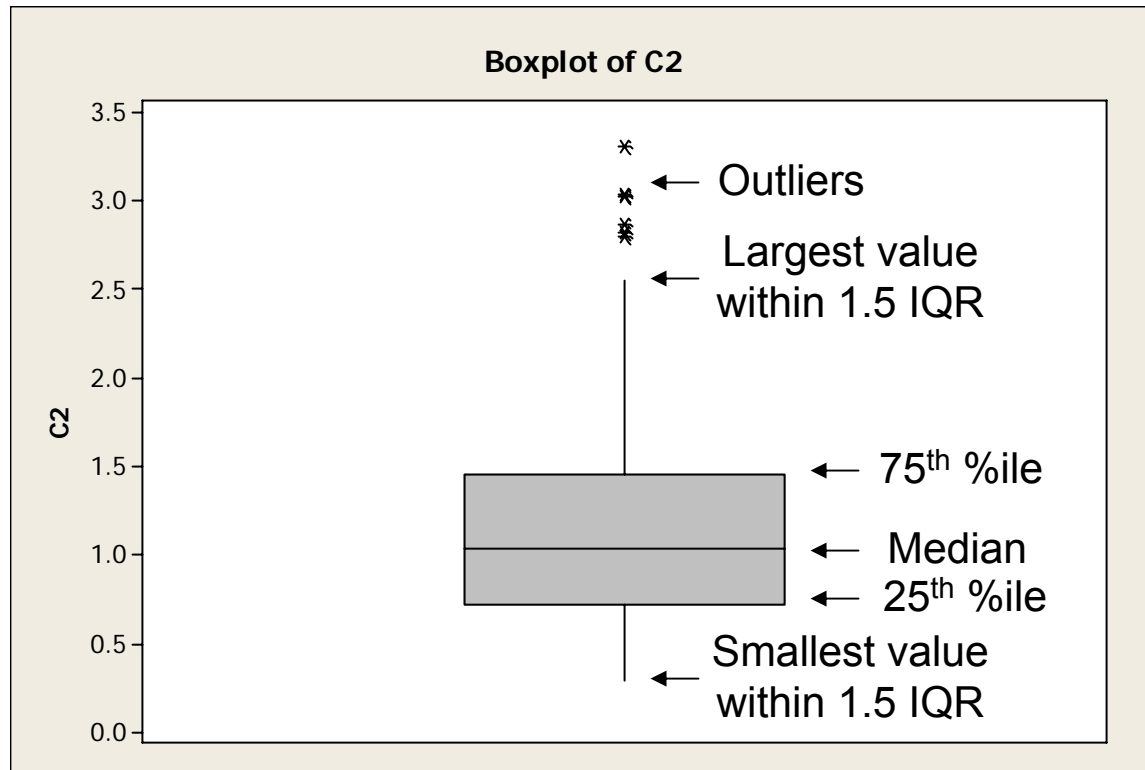
Range =  $16 - 1 = 15$ .

Interquartile range =  $4 - 2 = 2$ .

Like the median, the interquartile range is robust to outliers!

# Box plots

- Make it easy to see the variability and skewness of a distribution, as well as any outliers (unexpected values).



# Measures of variability

- **Variance:** the average squared deviation from the mean.
- **Standard deviation:** the square root of the variance.

$$\text{Sample variance } s^2 = \frac{\sum (x_i - \bar{x})^2}{n - 1}$$

$(n - 1)$  is used in the denominator instead of  $n$ .

This makes the sample variance  $s^2$  an unbiased estimator of the population variance  $\sigma^2$ .

$$\text{Sample standard deviation } s = \sqrt{s^2}$$

Example dataset: 1, 1, 2, 2, 2, 3, 4, 4, 5, 16

↓ Mean = 4

Deviations: -3, -3, -2, -2, -2, -1, 0, 0, 1, 12

Squared deviations: 9, 9, 4, 4, 4, 1, 0, 0, 1, 144

$$\text{Sample variance: } s^2 = (9 + 9 + 4 + 4 + 4 + 1 + 0 + 0 + 1 + 144) / (10 - 1) = \frac{176}{9}$$

$$\text{Sample standard deviation: } s = \sqrt{\frac{176}{9}} \approx 4.42$$

# Why measures of variability?

- Measures of the center tell us about our expectation (e.g. expected profit or loss).
- Measures of variability characterize our risk or uncertainty about this expectation.

Scenario 1: You are offered \$5000.

Expected profit? Risk? Would you take this offer?

Scenario 2: You are offered a gamble on the flip of a fair coin.  
If the coin comes up heads, you win \$50K, otherwise you lose \$40K.

Expected profit? Risk? Would you take this offer?

# The empirical rule

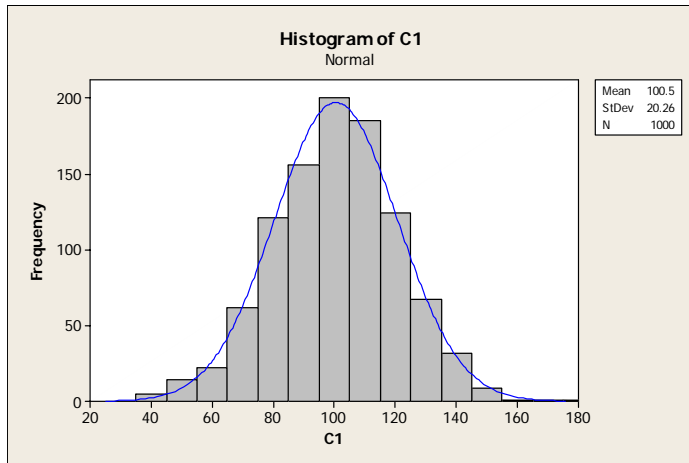
- For symmetric, unimodal (“mound-shaped”) distributions:
  - Approximately 68% of the measurements will fall within 1 standard deviation of the mean.
  - Approximately 95% of the measurements will fall within 2 standard deviations of the mean.
  - Approximately 99.7% of the measurements will fall within 3 standard deviations of the mean.
- This rule is useful for:
  - Identifying outliers (erroneous data, unusual events)
  - Calibrating the likelihood of success.
  - “Guesstimating” the standard deviation.

Example: mean height of trees = 30 feet, standard deviation = 10 feet

How likely are we to see a tree taller than 40 feet?

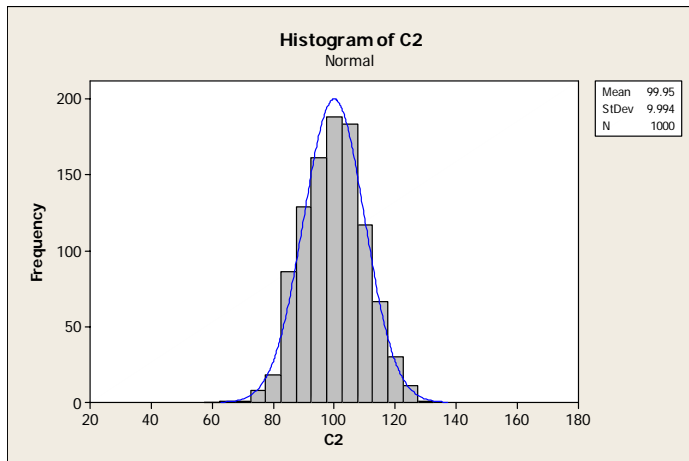
How likely are we to see a tree taller than 60 feet?

# Examples of the empirical rule



1000 data points generated from  $N(100,20)$

68% of the data should be between 80 and 120  
95% of the data should be between 60 and 140  
Almost all of the data should be between 40 and 160



1000 data points generated from  $N(100,10)$

68% of the data should be between 90 and 110  
95% of the data should be between 80 and 120  
Almost all of the data should be between 70 and 130

# Using Minitab

- Creating and listing data (p. 33-37)
- Graphing data (p. 134)
- Computing numerical descriptive statistics (p. 134-136)
- Generating a random sample (p. 203-204)

# Why study probability?

- Basis for statistical inference:
  - Margin of error on opinion poll is +/- 4%.
  - Difference between test scores is significant at 5% level.
- Key element of business:
  - Expected profit, risk, uncertainty, etc.
- Key element of operations management :
  - Setting inventory level, delivery cycle, response time.
- Our intuitions about probabilities are terrible!

“98% of individuals who do not make a return visit to a web site are first-time visitors.”

“98% of first-time visitors will not make a return visit to a web site.”

# Basic definitions

- **Probability of A:** a number  $P(A)$  between zero and one, indicating the likelihood of event A.
  - $P(\text{coin flip lands on heads}) = \frac{1}{2}$
  - $P(\text{it will rain tomorrow}) = 0.8$
- Interpreting probability as **relative frequency**:

$$P(A) = \lim_{n \rightarrow \infty} \frac{\text{\# of times event A occurs in n trials}}{n}$$

- Probabilities can be **objective** or **subjective**.
- **Complement** of event A: the event that A does not occur, usually denoted by  $\sim A$ ,  $A^C$ ,  $A'$ , or  $\bar{A}$ .
  - Important rule:  $P(\sim A) = 1 - P(A)$ .

# Combining probabilities

- Given two events A and B, the probability of both events occurring simultaneously is denoted by  $P(A \cap B)$ , i.e. the “probability of A and B.”
- The probability of at least one of the two events occurring is denoted by  $P(A \cup B)$ , i.e. the “probability of A or B.”
- Important rule:  $P(A \cup B) = P(A) + P(B) - P(A \cap B)$ 
  - Example:  $x =$  roll of a six-sided die.  $P(\{x \text{ is even}\} \cup \{x \geq 3\})$
- Mutually exclusive events:  $P(A \cap B) = 0$ .
  - For mutually exclusive events,  $P(A \cup B) = P(A) + P(B)$ .
  - Example:  $x =$  roll of a six-sided die.  $A = \{x \text{ is even}\}$ ,  $B = \{x = 1\}$ .
  - Example: A and  $\sim A$  are mutually exclusive and exhaustive.

$$P(A \cap \sim A) = 0$$

$$P(A \cup \sim A) = 1$$

# Conditional probabilities

- Given that an event B has occurred, the probability that event A has also occurred is denoted by  $P(A | B)$ , i.e. the “probability of A given B.”
  - Example:  $x =$  roll of a six-sided die.  $P(\{x \text{ is even}\} | \{x \leq 5\})$
- Important rule:  $P(A | B) = P(A \cap B) / P(B)$ .
  - Note that  $P(A | B) \neq P(B | A)$
  - Example:  $x =$  roll of a six-sided die.  $P(\{x \leq 5\} | \{x \text{ is even}\})$
- Another way to express this rule:  
$$P(A \cap B) = P(A | B) P(B) = P(B | A) P(A)$$
- Given mutually exclusive and exhaustive events  $B_1..B_n$ :  
$$P(A) = P(A \cap B_1) + P(A \cap B_2) + \dots + P(A \cap B_n)$$
$$= P(A | B_1) P(B_1) + P(A | B_2) P(B_2) + \dots + P(A | B_n) P(B_n).$$

Example: There are three coins in a box: one fair coin, one two-headed coin, and one biased coin with  $P(\text{heads}) = 2/3$ . If you draw one coin at random and flip it, what is the probability that it lands on heads?

# Independent events

- Two events  $A$  and  $B$  are said to be independent if:  
 $P(A | B) = P(A | \sim B) = P(A)$ , and  $P(B | A) = P(B | \sim A) = P(B)$ .
- In other words, two events are independent if the occurrence (or non-occurrence) of one event does not change the probability that the other will occur.
- Independent or dependent?
  - Example 1:  $A$  = heads on first toss of a fair coin,  $B$  = tails on second toss of that coin.
  - Example 2:  $A$  = individual knows Java programming,  $B$  = that individual is an engineer.
  - Example 3:  $A$  = heads on first toss of a fair coin,  $B$  = tails on first toss of that coin.
- If  $A$  and  $B$  are independent:  
 $P(A \cap B) = P(A | B) P(B) = P(A) P(B)$ .
- More generally, for independent events  $A_1 \dots A_n$ :  
 $P(A_1 \cap \dots \cap A_n) = P(A_1) P(A_2) \dots P(A_n)$ .

# Bayes' Theorem

- A way of figuring out a conditional probability  $P(A | B)$  if we have the opposite conditional probability,  $P(B | A)$ .
- In fact, we have to know the probabilities  $P(B | A)$  and  $P(B | \sim A)$ , as well as the “prior probability”  $P(A)$ .

$$P(A | B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A \cap B)}{P(A \cap B) + P(\sim A \cap B)} = \frac{P(B | A)P(A)}{P(B | A)P(A) + P(B | \sim A)P(\sim A)}$$

- More generally, given mutually exclusive and exhaustive events  $A_1 \dots A_n$ :

$$P(A_i | B) = \frac{P(A_i \cap B)}{P(B)} = \frac{P(B | A_i)P(A_i)}{P(B | A_1)P(A_1) + \dots + P(B | A_n)P(A_n)}$$

Example: There are three coins in a box: one fair coin, one two-headed coin, and one biased coin with  $P(\text{heads}) = 2/3$ . You draw one coin at random and flip it: it lands on heads. What is the probability that it is the fair coin?

# Random variables

- **Sample space:** the set of all possible outcomes of a statistical experiment.
  - Flipping three coins: HHH, HHT, ..., TTT
- **Random variable:** a variable that assigns a numerical value to each possible outcome.
  - Number of heads flipped: 3 if HHH, 2 if HHT, etc.
- Random variables can be **discrete** or **continuous**:
  - **Discrete variable** can take a countable number of values (e.g. number of heads flipped = 0, 1, 2, or 3).
  - **Continuous variable** can take an uncountable number of values (e.g. height, weight, response time).

# Discrete random variables

- **Probability mass function  $p(x)$**  specifies the probability associated with each possible value of the discrete random variable  $x$ .
  - Example:  $x$  = number of heads in three coin flips.

$p(0) = 1/8$	{TTT}
$p(1) = 3/8$	{TTH, THT, HTT}
$p(2) = 3/8$	{THH, HTH, HHT}
$p(3) = 1/8$	{HHH}
- We must have  $p(x) \geq 0$  for all  $x$ , and  $\sum p(x) = 1$ .
- **Mean (or expected value)**:  $\mu = \sum x p(x)$ .
- **Variance**:  $\sigma^2 = \sum (x - \mu)^2 p(x)$ .
- **Standard deviation**:  $\sigma = \sqrt{\sigma^2}$

What are the mean and standard deviation of  $x$  for the coin flip example?

# Sampling of random variables

- Let us assume that we perform the “three coin flip” experiment 80 times, and count the number of heads  $x$  for each experiment:
  - We expect: 10 { $x=0$ }, 30 { $x=1$ }, 30 { $x=2$ }, 10 { $x=3$ }.  
(Mean = 1.5, Variance = 0.75)
  - First trial: 12 { $x=0$ }, 22 { $x=1$ }, 31 { $x=2$ }, 15 { $x=3$ }.  
(Mean = 1.61, Variance = 0.92)
  - Second trial: 12 { $x=0$ }, 27 { $x=1$ }, 32 { $x=2$ }, 9 { $x=3$ }  
(Mean = 1.47, Variance = 0.78)
- Notice that the sample proportions are close, but not equal, to the expected proportions  $p(x)$ .
- As the number of trials increases, the sample proportions will converge to their expectations, as will the sample mean and sample variance.

“Law of Large Numbers”

# A practice problem

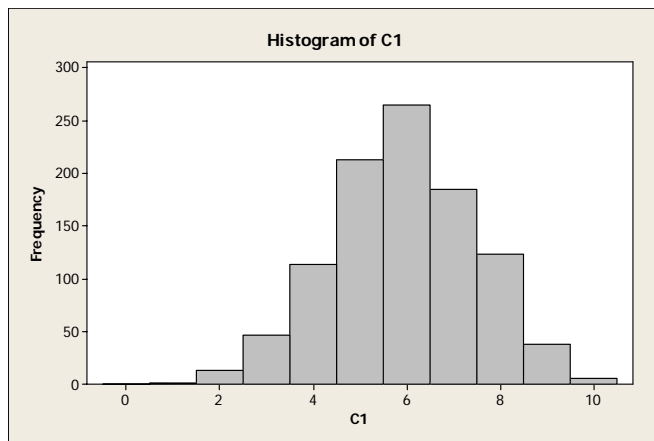
- An insurance company sells hurricane damage insurance to a Florida homeowner for \$1,000/year. In a given year, there is a 95% chance of no damage, 4% chance of minor (\$20,000) damage, and a 1% chance of major (\$80,000) damage.
  - Let  $x$  = the insurance company's profit. What is  $p(x)$ ?  
 $p(1,000) = 0.95$ ,  $p(-19,000) = 0.04$ ,  $p(-79,000) = 0.01$ .
  - What is the probability that the insurance company will make a profit in a given year?  
 $P(x > 0) = 95\%$ .
  - What is the company's expected yearly profit? Is this a profitable policy for the insurance company?  
 $0.95(\$1,000) + 0.04(-\$19,000) + 0.01(-\$79,000) = -\$600$ .  
Not profitable!

# The binomial distribution

- Given an experiment with probability  $p$  of success. Let random variable  $x$  denote the number of successes in  $n$  independent trials.
- Then  $x$  follows a **binomial distribution**,  $x \sim \text{Bin}(n,p)$ .

$$p(x) = \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x}, \text{ for } 0 \leq x \leq n$$

- For example, we have a weighted coin with  $P(\text{heads}) = 0.6$ . Let  $x =$  the number of heads in 10 trials.



$x \sim \text{Bin}(10,0.6)$

For  $x \sim \text{Bin}(n,p)$

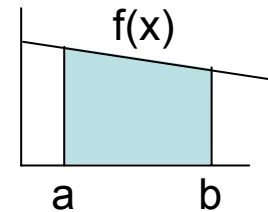
**Mean of  $x$ :  $\mu = np$ .**

**Variance of  $x$ :  $\sigma^2 = np(1-p)$**

# Continuous random variables

- **Probability density function  $f(x)$**  specifies the probability associated with each range of the continuous random variable  $x$ :

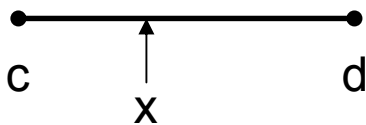
$$P(a \leq x \leq b) = \int_a^b f(x) dx \quad \leftarrow \text{Area under the curve } f(x), \text{ from } a \text{ to } b$$



- We must have  $f(x) \geq 0$  for all  $x$ , and  $\int f(x) dx = 1$ .
- **Mean (or expected value):**  $\mu = \int x f(x) dx$
- **Variance:**  $\sigma^2 = \int (x - \mu)^2 f(x) dx$
- **Standard deviation:**  $\sigma = \sqrt{\sigma^2}$

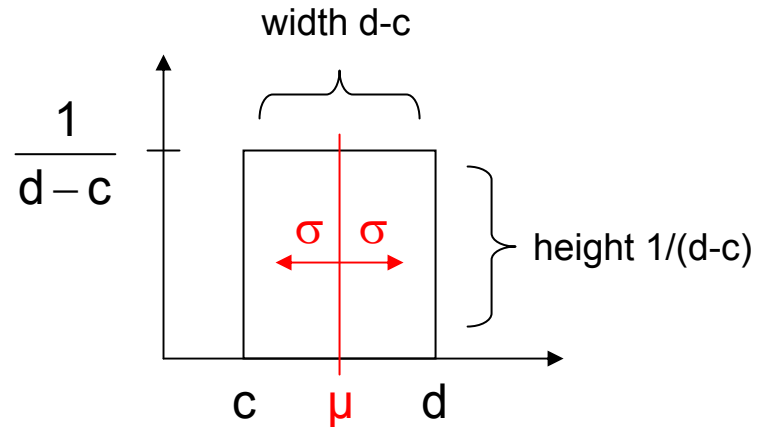
# The uniform distribution

- Choose a point on the interval  $[c,d]$ , where each point on the interval is equally likely.



$x \sim \text{Uniform}(c,d)$

$$f(x) = \begin{cases} \frac{1}{d-c} & \text{if } c \leq x \leq d \\ 0 & \text{otherwise} \end{cases}$$



**Mean:**  $\mu = (c + d) / 2$

**Variance:**  $\sigma^2 = (d - c)^2 / 12$

**Std. dev.:**  $\sigma = (d - c) / \sqrt{12}$

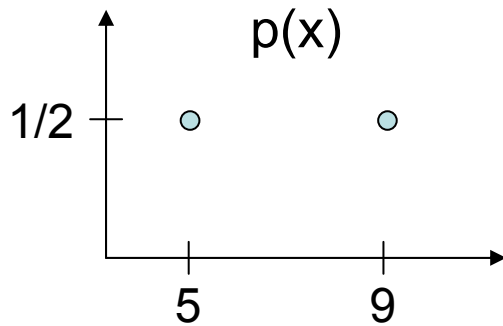
Example: if product weights are uniformly distributed on  $[1, 1.5]$ , what is the probability that a product will have weight  $> 1.2$ ?

# Comparison of discrete and continuous random variables

$x \sim \text{Endpoints}(5,9)$



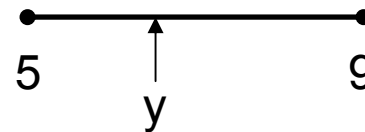
Probability mass function



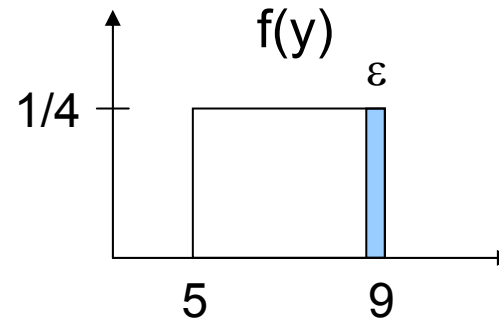
Sum of values = 1

$$\Pr(x = 5) = \Pr(x = 9) = \frac{1}{2}.$$

$y \sim \text{Uniform}(5,9)$



Probability density function



Area under curve = 1

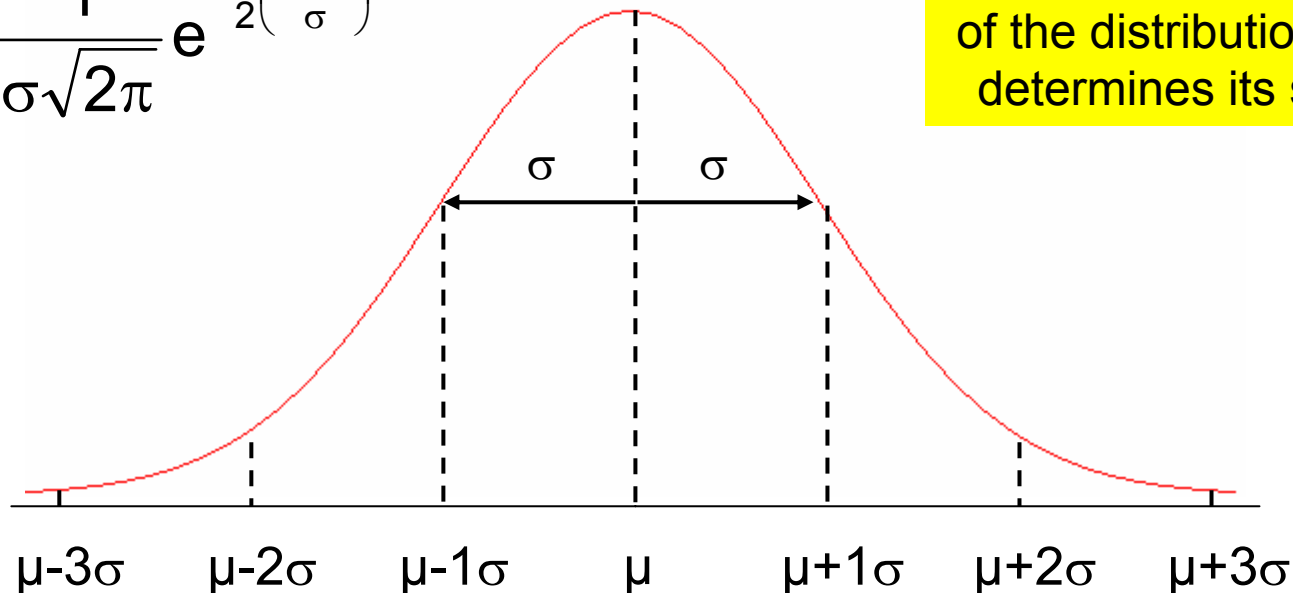
$$\Pr(9 - \epsilon \leq x \leq 9) = \epsilon / 4.$$

What are  $\mu$  and  $\sigma$  for each distribution?

# The Normal distribution

- The most important distribution for statistical inference!
  - Many real-world distributions are approximately normal.
- Also called “Gaussian distribution” or “bell curve”.
- A symmetric, unimodal distribution  $N(\mu, \sigma)$ , determined by its mean  $\mu$  and standard deviation  $\sigma$ :

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

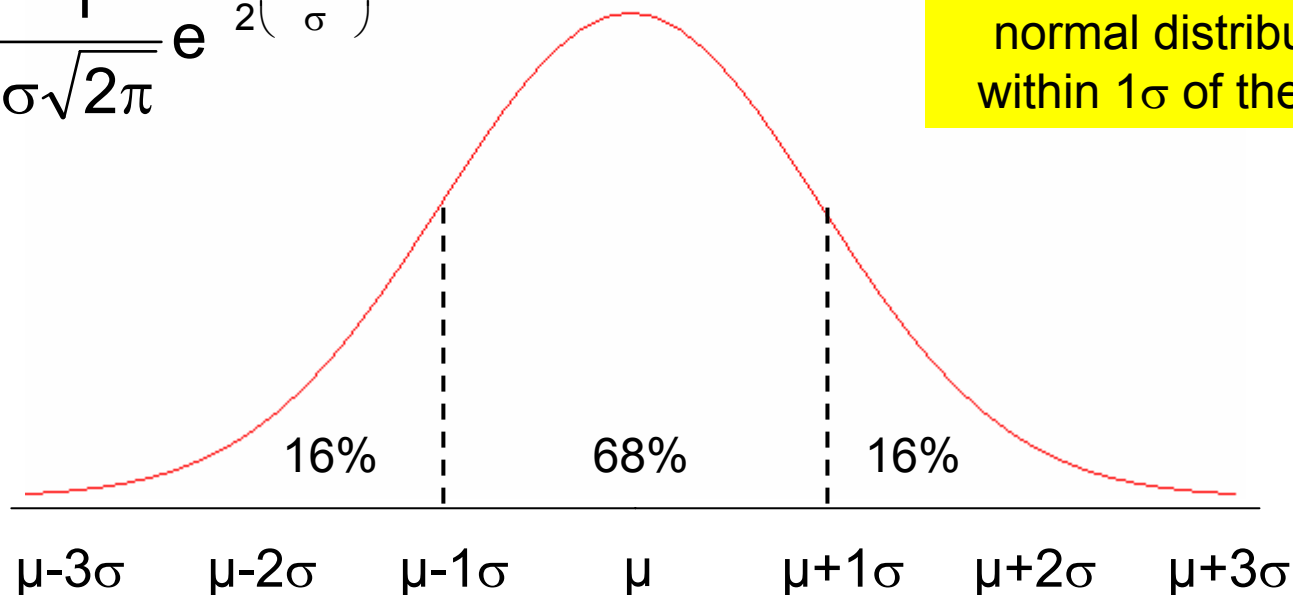


$\mu$  determines the center of the distribution, and  $\sigma$  determines its spread.

# The Normal distribution

- The most important distribution for statistical inference!
  - Many real-world distributions are approximately normal.
- Also called “Gaussian distribution” or “bell curve”.
- A symmetric, unimodal distribution  $N(\mu, \sigma)$ , determined by its mean  $\mu$  and standard deviation  $\sigma$ :

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$



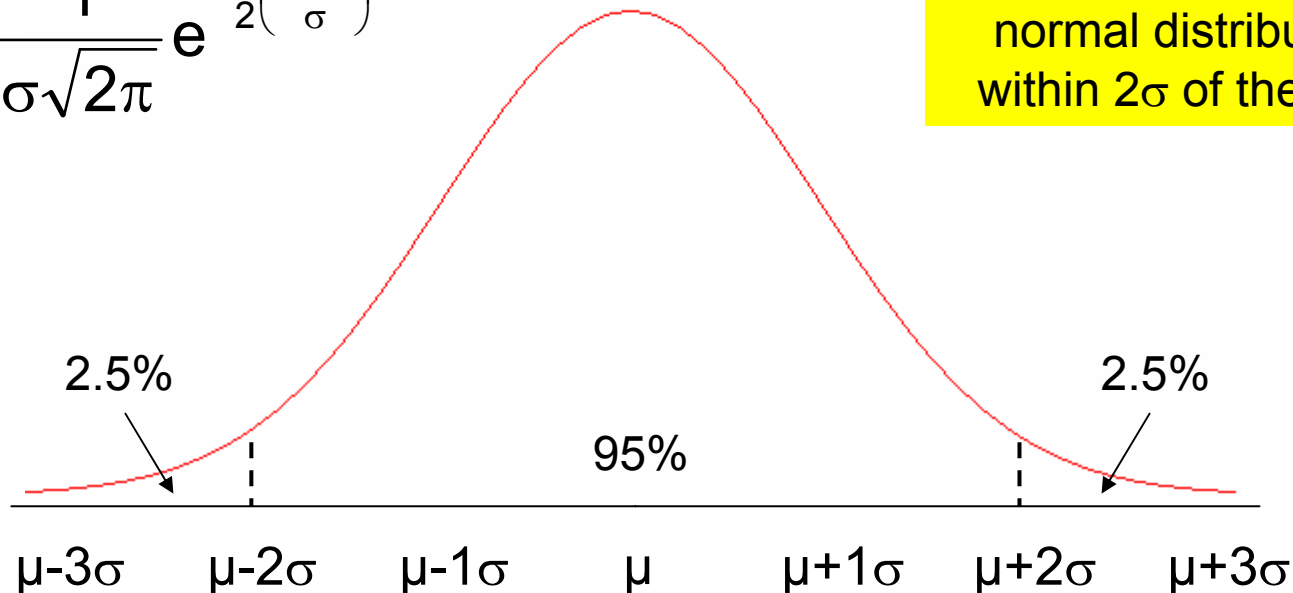
~68% of the area of the normal distribution is within  $1\sigma$  of the mean.

# The Normal distribution

- The most important distribution for statistical inference!
  - Many real-world distributions are approximately normal.
- Also called “Gaussian distribution” or “bell curve”.
- A symmetric, unimodal distribution  $N(\mu, \sigma)$ , determined by its mean  $\mu$  and standard deviation  $\sigma$ :

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

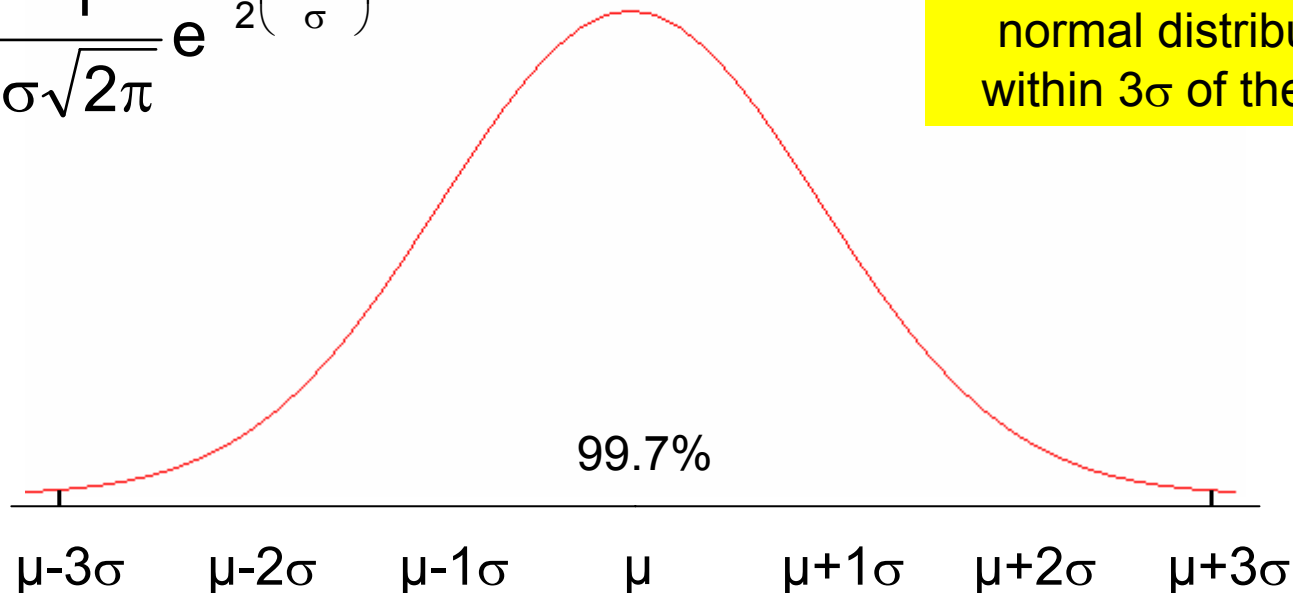
~95% of the area of the normal distribution is within  $2\sigma$  of the mean.



# The Normal distribution

- The most important distribution for statistical inference!
  - Many real-world distributions are approximately normal.
- Also called “Gaussian distribution” or “bell curve”.
- A symmetric, unimodal distribution  $N(\mu, \sigma)$ , determined by its mean  $\mu$  and standard deviation  $\sigma$ :

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

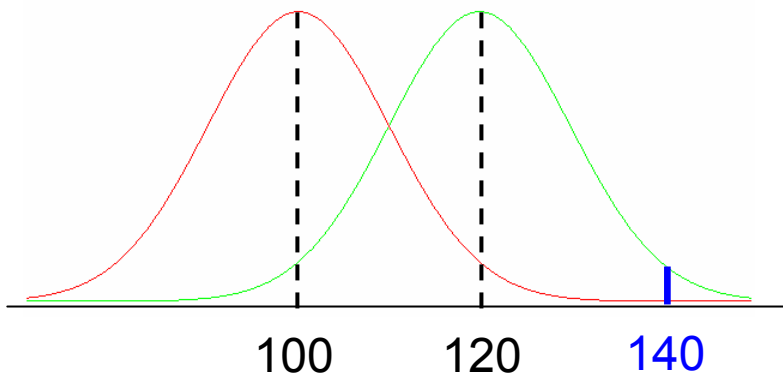


~99.7% of the area of the normal distribution is within  $3\sigma$  of the mean.

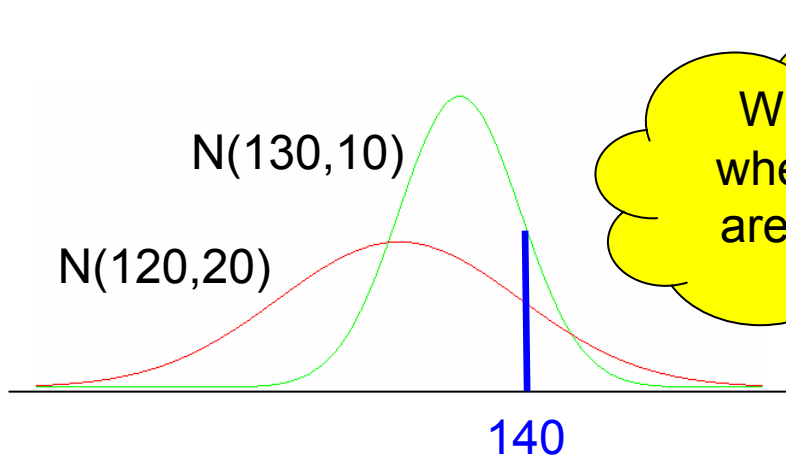
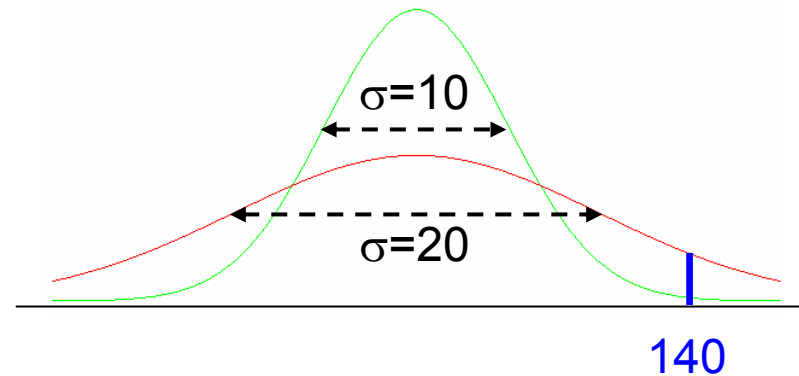
# Computing normal probabilities

- Normal probabilities depend both on  $\mu$  and  $\sigma$ .
  - Example: which has higher probability of  $x > 140$ ?

Same  $\sigma=10$ , different  $\mu$



Same  $\mu=110$ , different  $\sigma$

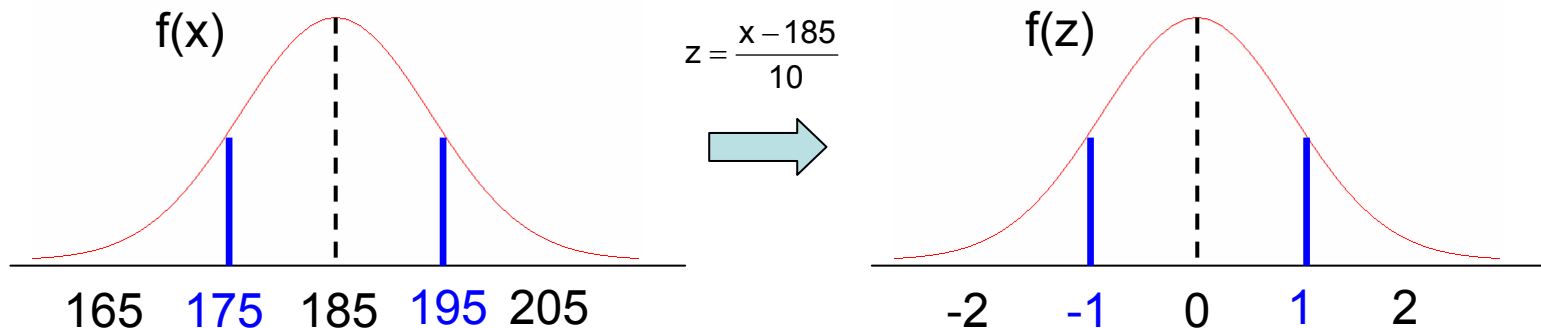


What about  
when  $\mu$  and  $\sigma$   
are different?

Solution: transform  
each distribution  
using the **z-score**!

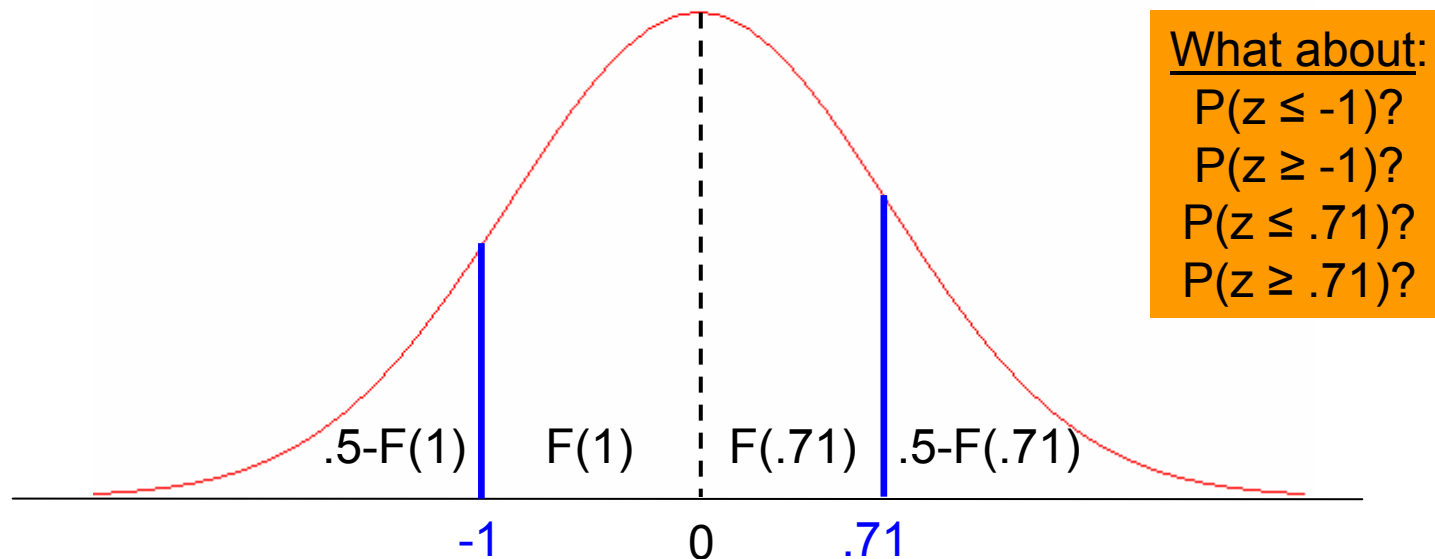
# Computing z-scores

- If  $x$  is distributed according to  $N(\mu, \sigma)$ , then  $z = \frac{x - \mu}{\sigma}$  will be distributed according to the **standard normal distribution**,  $N(0, 1)$ .
  - The **z-score** ( $z$ ) is the number of standard deviations ( $\sigma$ ) that the original measurement ( $x$ ) is from the mean ( $\mu$ ).
  - Example: man's weight  $x \sim N(185, 10)$ .  
 $P(175 \leq x \leq 195) = P(-1 \leq z \leq 1) \approx 68\%$ .



# Using a table of normal curve areas

- Once we have converted to z-scores, how do we compute more general probabilities, e.g.  $P(-1 \leq z \leq .71)$ ?
- Answer: use a table of normal curve areas (or Minitab).
  - The table gives  $F(z_0) = P(0 \leq z \leq |z_0|)$ .
  - We can use these values to compute any desired probability.
- Example:  $P(-1 \leq z \leq .71) = F(1) + F(.71) = .3413 + .2611 = .6024$



# A practice problem

- Let us assume that men's weights are normally distributed with  $\mu = 185$  and  $\sigma = 20$ , while women's weights are normally distributed with  $\mu = 150$  and  $\sigma = 10$ . Are men or women more likely to have weight between 160 and 170?

1<sup>st</sup> step: Convert to z-scores

$$\text{Men: } P(160 < x < 170) = P(-1.25 < z < -.75)$$

$$\text{Women: } P(160 < x < 170) = P(1 < z < 2)$$

2<sup>nd</sup> step: Compute probabilities

$$\text{Men: } P(-1.25 < z < -.75) = F(1.25) - F(.75) = .3944 - .2734 = .1210$$

$$\text{Women: } P(1 < z < 2) = F(2) - F(1) = .4772 - .3413 = .1359$$

# An “inverse” problem

- Large employers regularly use skill tests to evaluate potential employees. Suppose a test of programming proficiency has a mean score of 60% and standard deviation of 10%. If the employer only wants to hire the most proficient 20% of applicants, what is the minimum test score they should set?

1<sup>st</sup> step: Compute the necessary range of z-scores

$$P(z > z_0) = 0.2$$

$$P(0 < z < z_0) = 0.5 - 0.2 = 0.3$$

$$z_0 = F^{-1}(0.3) \approx 0.84$$

2<sup>nd</sup> step: Compute the necessary range of values

$$z > 0.84$$

$$x > 60\% + 0.84(10\%) \rightarrow x > 68.4\%$$

What if the employer wants to avoid hiring the bottom 20% of applicants?

# Why the normal distribution?

- Central Limit Theorem: averages are approximately normally distributed.
  - More samples = closer to a normal distribution.
  - More samples = lower variance.
- Other probability distributions (e.g. binomial) can be expressed as a sum, and thus are also approximately normally distributed.
- These properties will be very useful for inference (confidence intervals and hypothesis testing), as we will discuss in Module II.

# Parameters and sample statistics

- If we know the probability distribution of a random variable, we can compute its mean  $\mu$ , standard deviation  $\sigma$ , and associated probabilities.
  - “The average response time in minutes for a network outage is normally distributed with  $\mu = 47$ ,  $\sigma = 18$ .”
- What if we don't know the distribution, but only have samples from this distribution?
  - “For the last 5 network outages, response times were 43, 79, 21, 71, and 51 minutes ( $\bar{x} = 53$ ,  $s \approx 23$ ).”

What can we conclude about **population parameters**  $\mu$  and  $\sigma$ , using the **sample statistics**  $\bar{x}$  and  $s$ ?

# Parameters and sample statistics

- If we know the probability distribution of a random variable, we can compute its mean  $\mu$ , standard

The sample mean  $\bar{x}$  can be used as an estimate of the population mean  $\mu$ . But how good an estimate is it?

- Intuitively,  $\bar{x}$  will be a good estimate if the number of samples is large, and a poor estimate if the number of samples is small.
  - “For the last 5 network outages, response times were 43, 79, 21, 71, and 51 minutes ( $\bar{x} = 53$ ,  $s \approx 23$ ).”

What can we conclude about **population parameters**  $\mu$  and  $\sigma$ , using the **sample statistics**  $\bar{x}$  and  $s$ ?

# Sampling distributions

- A parameter such as  $\mu$  or  $\sigma$  describes some characteristic of a population. It is a fixed quantity that is calculated from all observations in the population.
- A sample statistic such as  $\bar{x}$  or  $s$  describes some characteristic of a sample. It is calculated only from those members of the population that are included in the sample.
- Since the value of a sample statistic will be different for each sample, a sample statistic is a random variable.
  - The probability distribution of this random variable is called its sampling distribution.

# Sampling distributions

- Example: You want to know the proportions of children and adults in a room.
- You observe only two of the five people in the room: let  $x$  be the proportion of children in the sample.
- If there are actually four adults and one child, what is the sampling distribution of  $x$ ?

$$p(0) = 6/10 \quad \{A_1A_2, A_1A_3, A_1A_4, A_2A_3, A_2A_4, A_3A_4\}$$

$$p(1/2) = 4/10 \quad \{A_1C, A_2C, A_3C, A_4C\}$$

$$\mu_x = 1/5$$

$$\sigma_x \approx .24$$

The sample statistic  $x$  is an unbiased estimate of the proportion of children in the population.

# Sampling distributions

- Example: You want to know the proportions of children and adults in a room.
- You observe only **four** of the five people in the room: let  $x$  be the proportion of children in the sample.
- If there are actually four adults and one child, what is the sampling distribution of  $x$ ?

$$p(0) = 1/5 \quad \{A_1A_2A_3A_4\}$$

$$p(1/4) = 4/5 \quad \{A_1A_2A_3C, A_1A_2A_4C, A_1A_3A_4C, A_2A_3A_4C\}$$

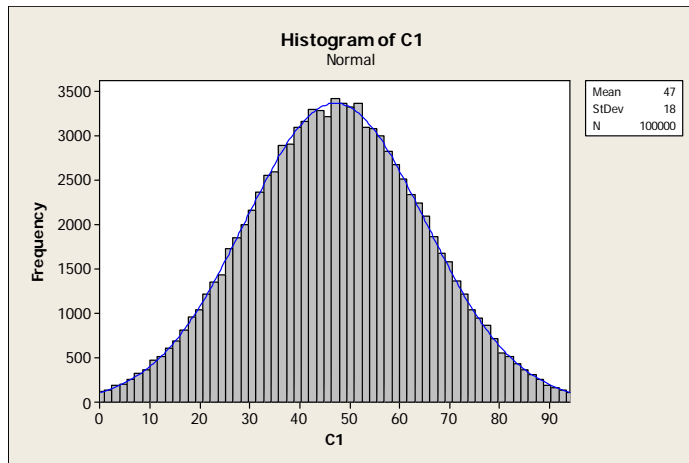
$$\mu_x = 1/5$$

$$\sigma_x = .10 \quad \text{— Larger sample size leads to a lower variance of the sampling distribution, i.e. better estimates!}$$

# Using $\bar{x}$ to estimate $\mu$

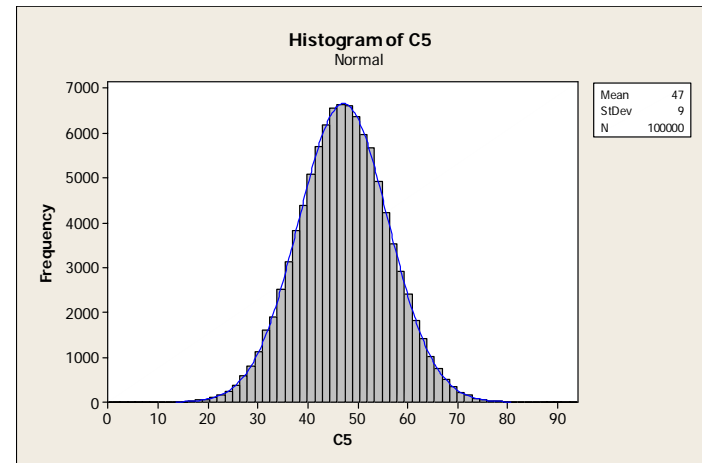
Let us assume that the population is normally distributed with  $\mu = 47$ ,  $\sigma = 18$ .

Here is a histogram of 100,000 samples drawn from the population.



Now consider drawing  $N = 4$  samples from the population and taking their mean,  $\bar{x}$ .

We repeat this experiment 100,000 times and form a histogram of the values of  $\bar{x}$ .

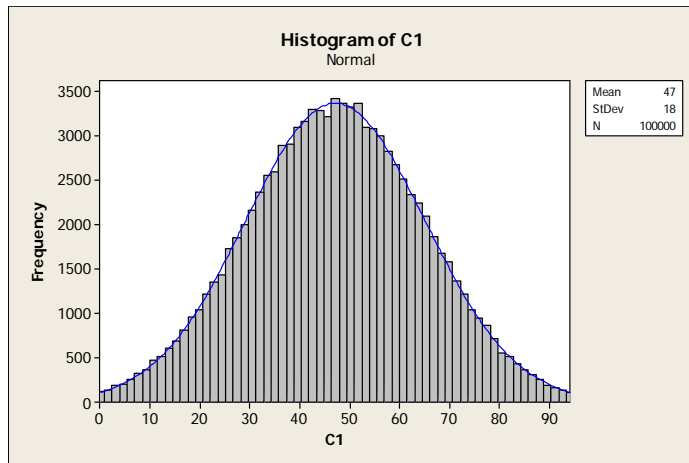


The sampling distribution of  $\bar{x}$  is normal, with mean  $\mu_{\bar{x}} = 47$  and standard deviation  $\sigma_{\bar{x}} = 9$ . Notice that the sample mean  $\bar{x}$  is an unbiased estimator of the population mean  $\mu$ . Additionally, the sample mean will be between 38 and 56 about 68% of the time.

# Using $\bar{x}$ to estimate $\mu$

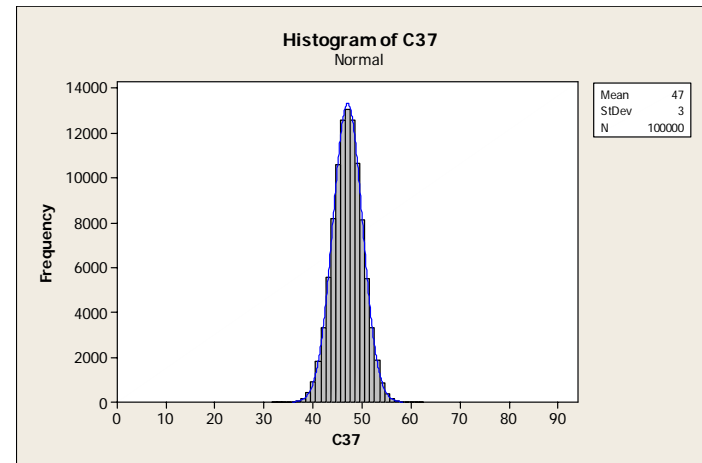
Let us assume that the population is normally distributed with  $\mu = 47$ ,  $\sigma = 18$ .

Here is a histogram of 100,000 samples drawn from the population.



Now consider drawing **N = 36** samples from the population and taking their mean,  $\bar{x}$ .

We repeat this experiment 100,000 times and form a histogram of the values of  $\bar{x}$ .

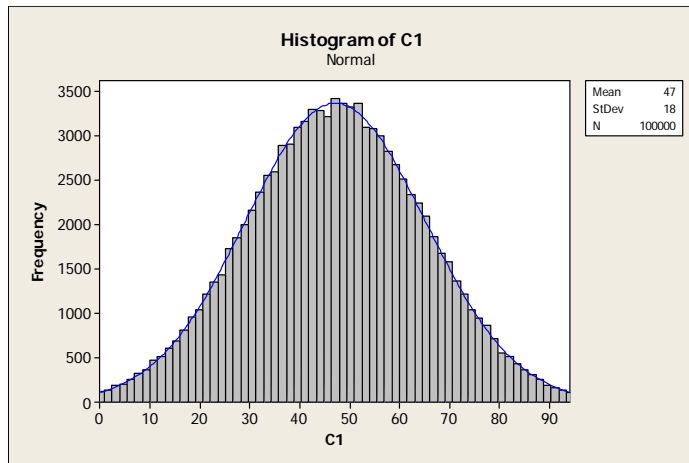


The sampling distribution of  $\bar{x}$  is normal, with mean  $\mu_{\bar{x}} = 47$  and standard deviation  $\sigma_{\bar{x}} = 3$ . Notice that the sample mean  $\bar{x}$  is an unbiased estimator of the population mean  $\mu$ . Additionally, the sample mean will be between **44** and **50** about 68% of the time.

# Using $\bar{x}$ to estimate $\mu$

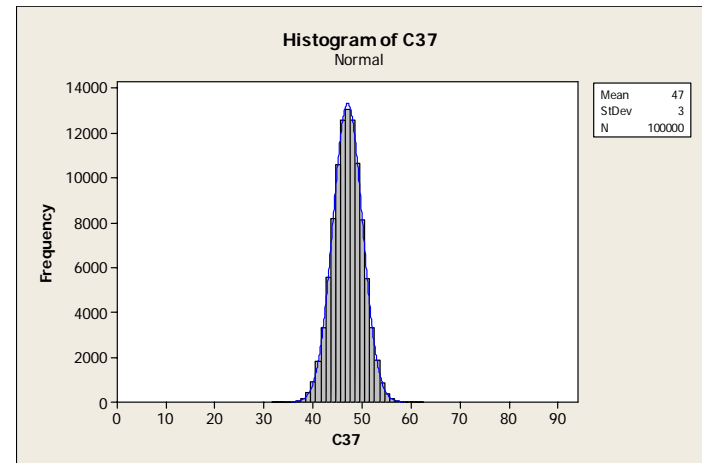
Let us assume that the population is normally distributed with  $\mu = 47$ ,  $\sigma = 18$ .

Here is a histogram of 100,000 samples drawn from the population.



Now consider drawing **N = 36** samples from the population and taking their mean,  $\bar{x}$ .

We repeat this experiment 100,000 times and form a histogram of the values of  $\bar{x}$ .

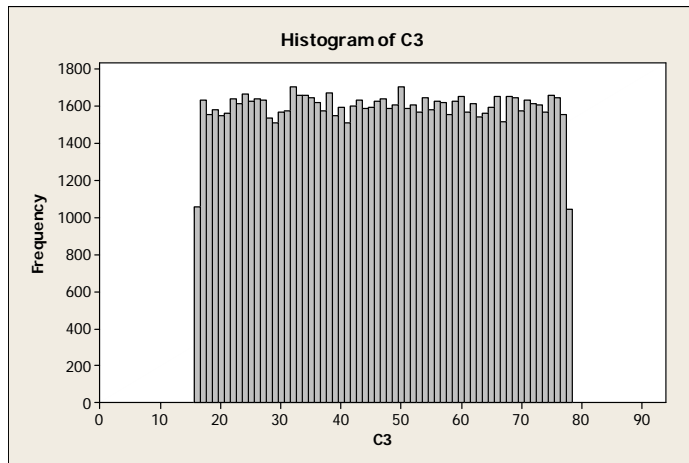


If the population is normally distributed with mean  $\mu$  and standard deviation  $\sigma$ , then the sample mean  $\bar{x}$  is also normally distributed, with mean  $\mu$  and standard deviation  $\sigma / \sqrt{N}$ .

# Using $\bar{x}$ to estimate $\mu$

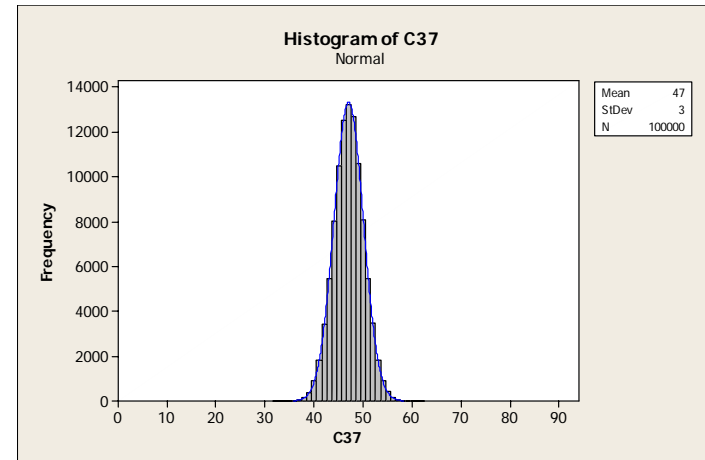
Let us assume that the population is **uniformly** distributed with  $\mu = 47$ ,  $\sigma = 18$ .

Here is a histogram of 100,000 samples drawn from the population.



Now consider drawing  $N = 36$  samples from the population and taking their mean,  $\bar{x}$ .

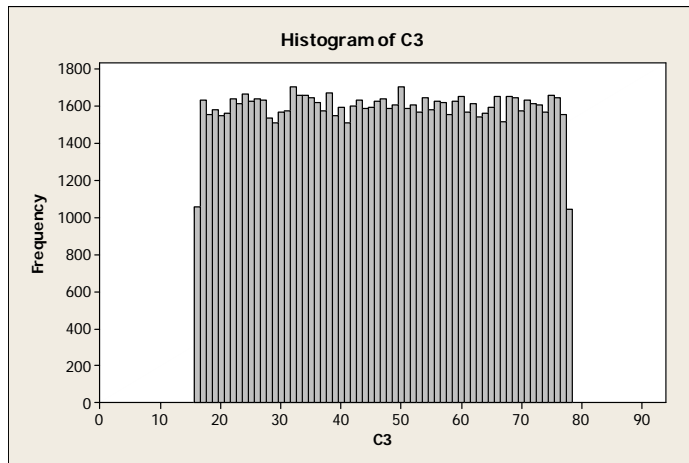
We repeat this experiment 100,000 times and form a histogram of the values of  $\bar{x}$ .



# Using $\bar{x}$ to estimate $\mu$

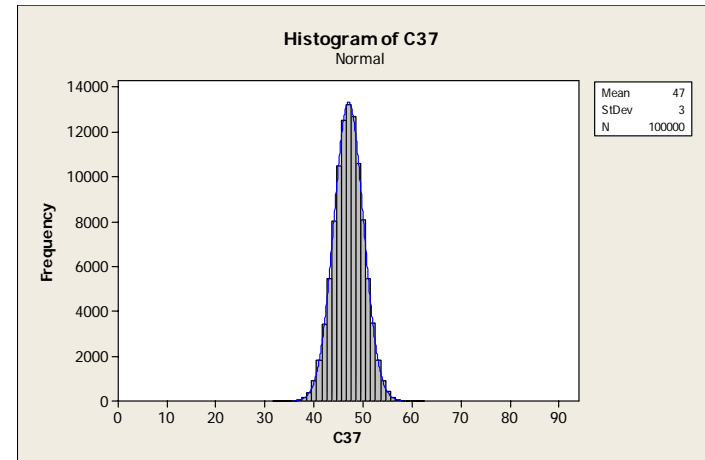
Let us assume that the population is **uniformly** distributed with  $\mu = 47$ ,  $\sigma = 18$ .

Here is a histogram of 100,000 samples drawn from the population.



Now consider drawing  $N = 36$  samples from the population and taking their mean,  $\bar{x}$ .

We repeat this experiment 100,000 times and form a histogram of the values of  $\bar{x}$ .



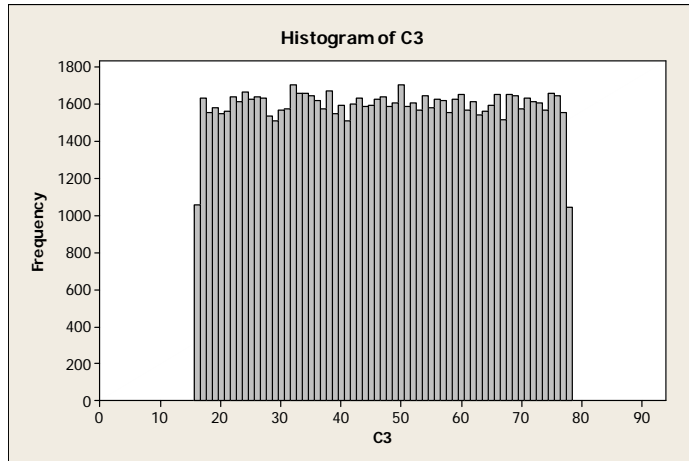
If the population has any distribution with mean  $\mu$  and standard deviation  $\sigma$ , **and if  $N \geq 30$** , then the sample mean  $\bar{x}$  is normally distributed, with mean  $\mu$  and standard deviation  $\sigma / \sqrt{N}$ .

This rule is called the Central Limit Theorem.

# What if N is too small?

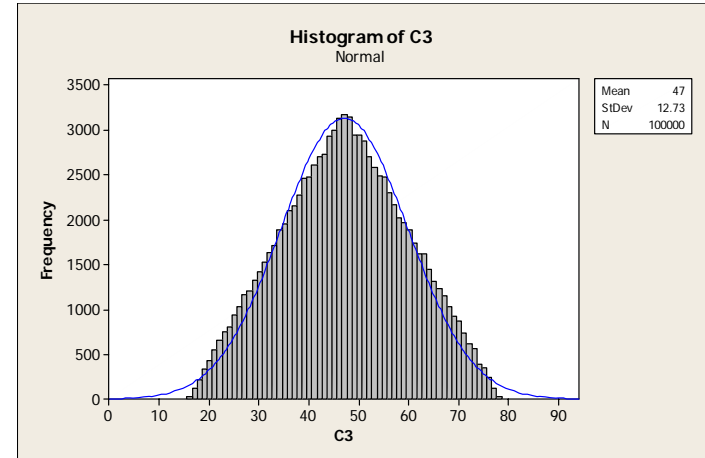
Let us assume that the population is uniformly distributed with  $\mu = 47$ ,  $\sigma = 18$ .

Here is a histogram of 100,000 samples drawn from the population.



Now consider drawing **N = 2** samples from the population and taking their mean,  $\bar{x}$ .

We repeat this experiment 100,000 times and form a histogram of the values of  $\bar{x}$ .



In general, the sample mean  $\bar{x}$  has mean  $\mu$  and standard deviation  $\sigma / \sqrt{N}$ , but it is only approximately normal for large N.

# The Central Limit Theorem

If the population has any distribution with mean  $\mu$  and standard deviation  $\sigma$ , **and if  $N \geq 30$** , then the sample mean  $\bar{x}$  is normally distributed, with mean  $\mu$  and standard deviation  $\sigma / \sqrt{N}$ .

Example problem: if the daily number of hits for your website follows some distribution with  $\mu = 1000$  and  $\sigma = 300$ , what is the probability that you will receive more than 39,600 hits in the next 36 days?

Given  $\mu = 1000$ ,  $\sigma = 300$ , and  $N = 36$ , we know that the sample mean  $\bar{x}$  is normally distributed with  $\mu_{\bar{x}} = 1000$  and  $\sigma_{\bar{x}} = 300 / \sqrt{36} = 50$ .

Then  $\Pr(\bar{x} > \frac{39,600}{36}) = \Pr(\bar{x} > 1100) = \Pr(z > \frac{1100 - 1000}{50}) = \Pr(z > 2)$ .  
Using the table of normal curve areas, we obtain  $.5 - .4772 = .0228$ .

Given  $\mu$  and  $\sigma$ , the Central Limit Theorem lets you reason about  $\bar{x}$ .

# The Central Limit Theorem

Example problem #2: An analyst for an internet consulting company is charged with collecting data on the performance of file sharing networks. A network is rated “satisfactory” if the average number of retries needed to gain entry is at most 1.

The analyst tests a site by attempting to gain entry 100 times. She finds a mean of 1.5 retries and a standard deviation of 1. Can she reliably conclude that the performance of the site is unsatisfactory?

Let us assume that  $\sigma \approx s = 1$ . Does a sample mean of  $\bar{x} = 1.5$ , computed from  $N = 100$  trials, seem consistent with the assumption that the population mean  $\mu$  is equal to 1?

If the population had  $\mu = 1$  and  $\sigma = 1$ , we would expect  $\bar{x}$  to be normally distributed with mean 1 and std. deviation  $1 / \sqrt{100} = 0.1$ .

Then  $\Pr(\bar{x} \geq 1.5) = \Pr(z \geq 5) \approx 0$ .

Given  $\bar{x}$  and  $s$ , the Central Limit Theorem lets you reason about  $\mu$ .