STA 321 Spring 2014

Lecture 10 *Thursday, Feb 20*

- **Normal Distribution**
- **z-Scores**

Homework 6: Due today!

Bernoulli Trial

- Suppose we have a single random experiment X with two outcomes: “success” and “failure.”
- Typically, we denote “success” by the value 1 and “failure” by the value 0.
- It is also customary to label the corresponding probabilities as:

  $P(\text{success}) = P(1) = p$ and

  $P(\text{failure}) = P(0) = 1 - p = q$

- Note: $p + q = 1$

Binomial Distribution I

- Suppose we perform several Bernoulli experiments and they are all independent of each other.
- Let's say we do *n* of them. The value *n* is the **number of trials.**
- We will label these *n* Bernoulli random variables in this manner: $X_1, X_2, \ldots, X_n$
- As before, we will assume that the probability of success in a single trial is *p*, and that this probability of success doesn't change from trial to trial.

Binomial Distribution II

- Now, we will build a new random variable X using all of these Bernoulli random variables:

$$X = X_1 + X_2 + \cdots + X_n = \sum_{i=1}^{n} X_i$$

- What are the possible outcomes of X?
- What is X counting?
- How can we find $P(X = x)$?

Binomial Distribution III

- We need a quick way to count the number of ways in which *k* successes can occur in *n* trials.
- Here's the formula to find this value:

$$\binom{n}{k} = {}_nC_k = \frac{n!}{k!(n-k)!}, \text{ where } n! = n \cdot (n-1) \cdot \cdots \cdot 3 \cdot 2 \cdot 1 \text{ and } 0! \equiv 1$$

- Note: ${}_nC_k$ *is read as "n* choose *k."*

Binomial Distribution IV

- Now, we can write the formula for the binomial distribution:
- The probability of observing *x* successes in *n* independent trials is

$$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}, \text{ for } x = 0, 1, \cdots, n$$

under the assumption that the probability of success in a single trial is *p.*

Using Binomial Probabilities

Note: Unlike generic random variables where we would have to be given the probability distribution or calculate it from a frequency distribution, here we can calculate it from a mathematical formula.

Helpful resources (besides your calculator):

- Excel:

Enter	Gives
=BINOMDIST(4,10,0.2,FALSE)	0.08808
=BINOMDIST(4,10,0.2,TRUE)	0.967207

Binomial Probabilities

We are choosing a random sample of $n = 7$ Lexington residents—our random variable, C = number of Centerpointe supporters in our sample. Suppose, $p = P$ (Centerpointe support) ≈ 0.3. Find the following probabilities:

a) $P(C = 2)$ http://stattrek.com/Tables/Binomial.aspx

b) $P(C < 2)$

c) $P(C \leq 2)$

d) $P(C \geq 2)$

e) $P(1 \leq C \leq 4)$

What is the *expected* number of Centerpointe supporters, μ_C?

The Normal (*Gaussian, Bell Curve*) Distribution

- Carl Friedrich Gauß (1777-1855), ***Gaussian Distribution***

- Normal distribution is perfectly ***symmetric*** and ***bell-shaped***
- Characterized by two parameters:

  mean μ and ***standard deviation σ***
- The ***68%-95%-99.7%*** rule applies "precisely"* to the normal distribution

*More precisely: 68.26895% - 95.44997% - 99.73002%


3D9
DEUTSCHE BUNDESBANK
Banknote
10
1777 1855 Carl Friedr. Gauß
10
ZEHN DEUTSCHE MARK
DN6833178D9

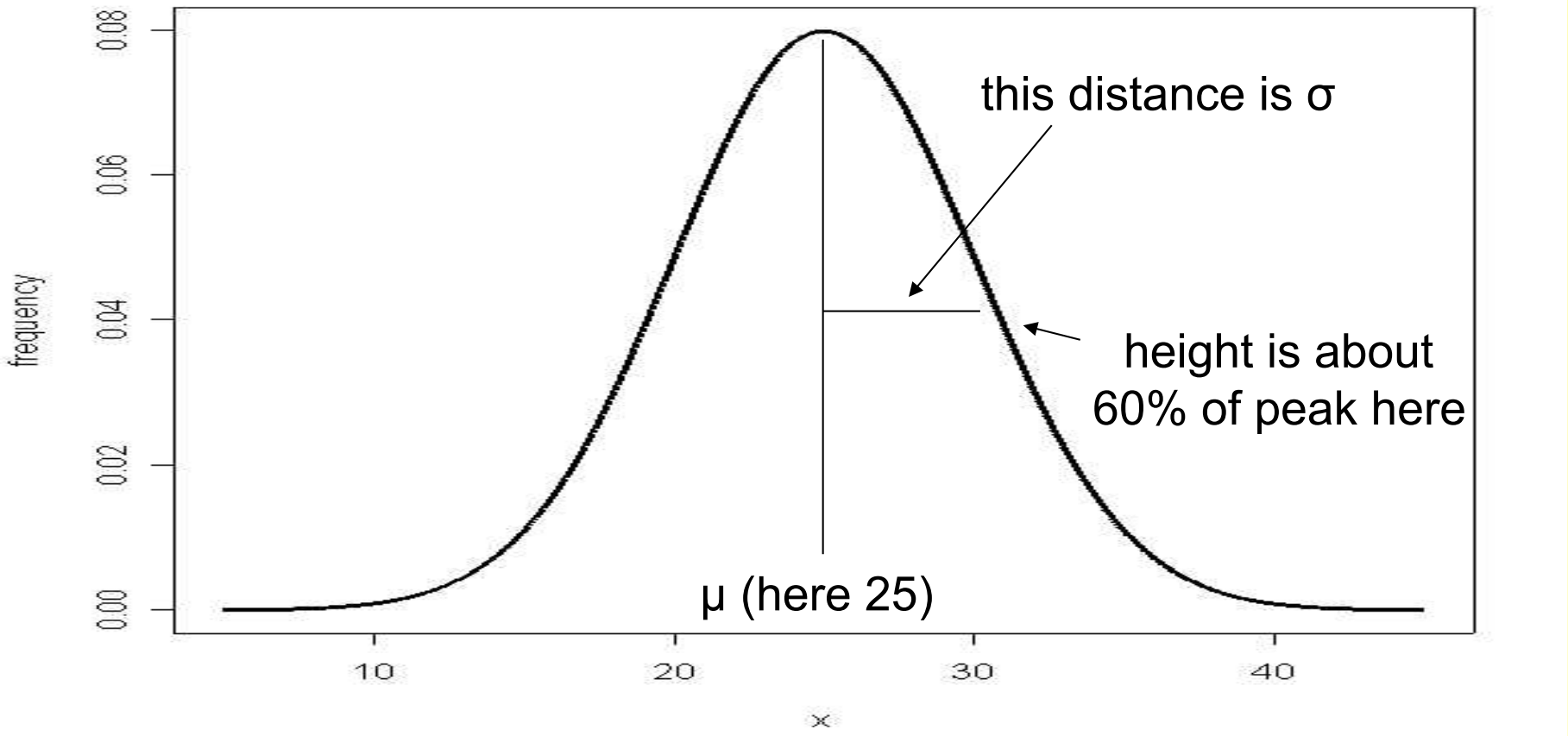
The Normal Distribution is Common

- Many real data follow a normal shape. For example
- 1) Many/most biometric measurements (heights, femur lengths, skull diameters, etc.)
- 2) Scores on many standardized exams (IQ tests, SAT, ACT) are forced into a normal shape before reporting
- 3) Microarray expression intensities (if you take the logarithm first)
- 4) Averages of measurements!

Mean and Standard Deviation

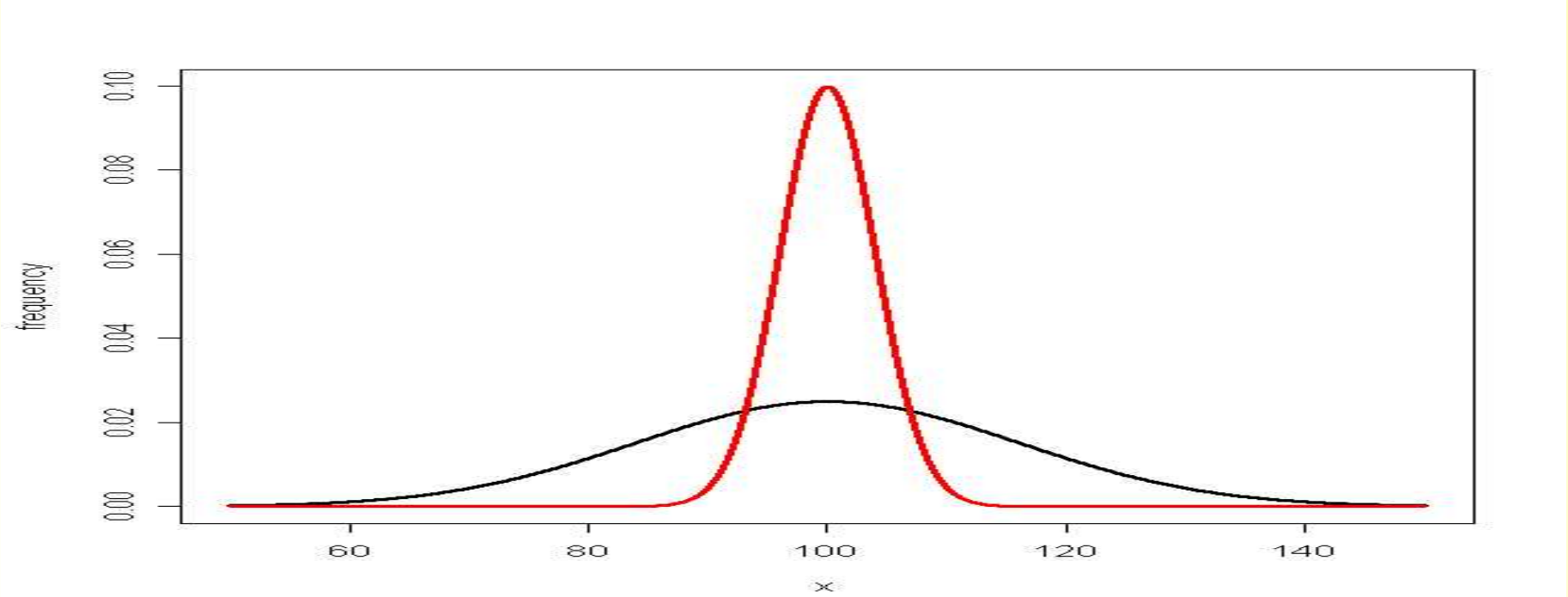
- Normal distributions are characterized by two numbers
- mean or “expected value” (corresponding to the peak)
- “standard deviation” (distance from mean to inflection point)
- Large standard deviations result in “spread out” normal distributions.
- Small standard deviations result in “strongly peaked” distributions.

Mean (μ) and Standard deviation (σ) for a normal distribution

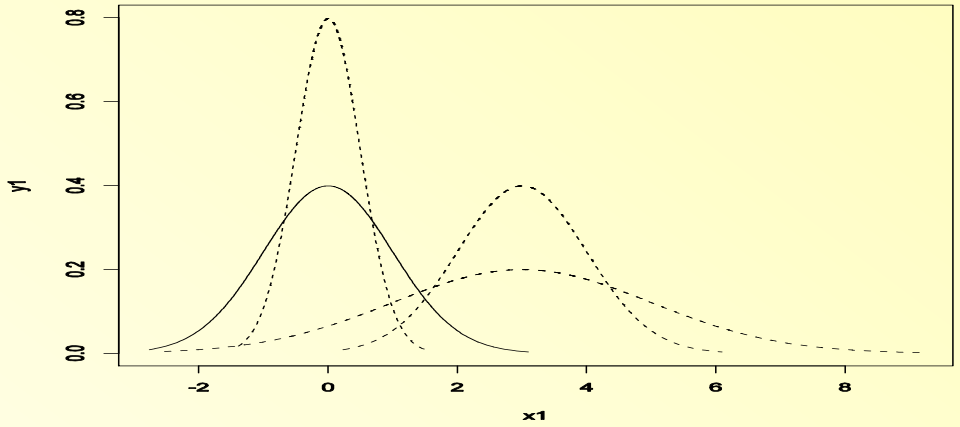


Two Normal Distributions, Corresponding to Different Standard Deviations

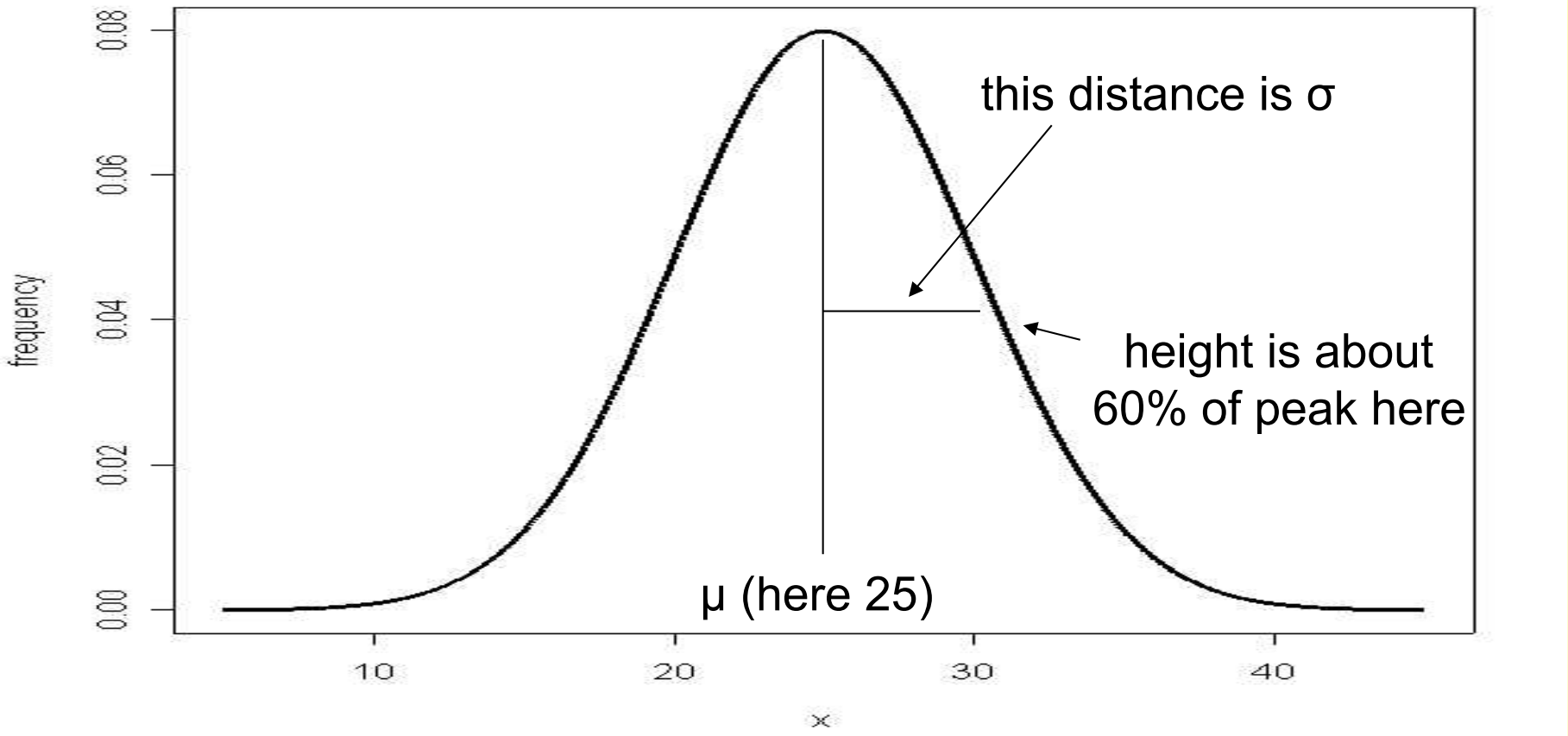
- Mean=100, std.dev = 16
- Mean=100, std.dev = 4



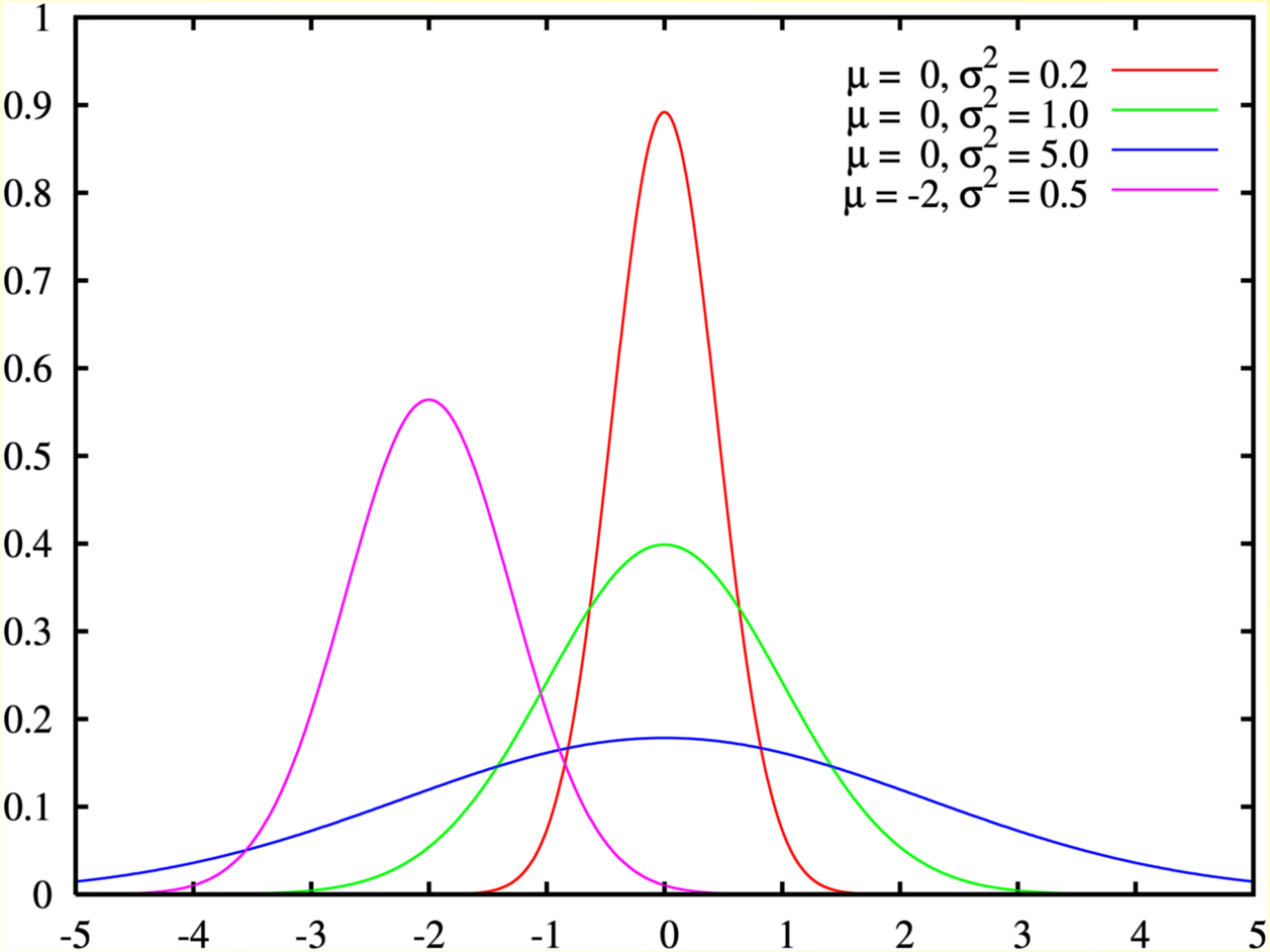
More Normal Distributions



Mean (μ) and Standard deviation (σ) for a normal distribution



More on normal distribution



Describing Normal Distributions

- Central location: mean μ (=median).
- Spread: standard deviation σ (interquartile range is about 4/3 σ)
- Shape: Normal distributions are symmetric and typically have few, if any, outliers.

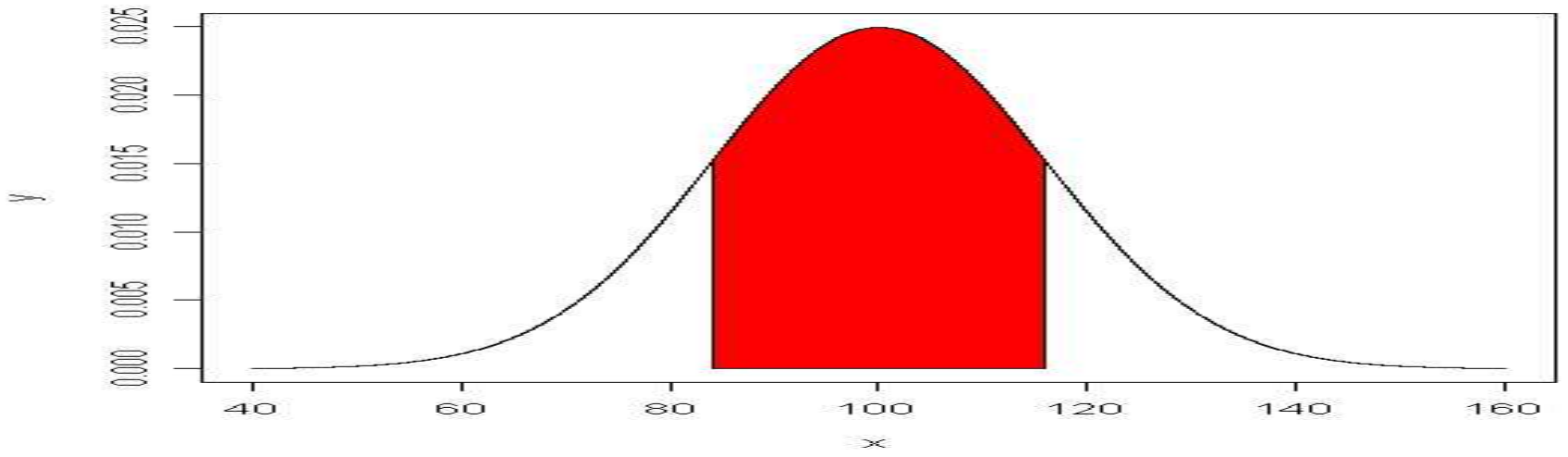
- If your data has a lot of outliers, but is otherwise symmetric and unimodal, it may have a “t” distribution (discussed later in class).

Probabilities from a Normal distribution

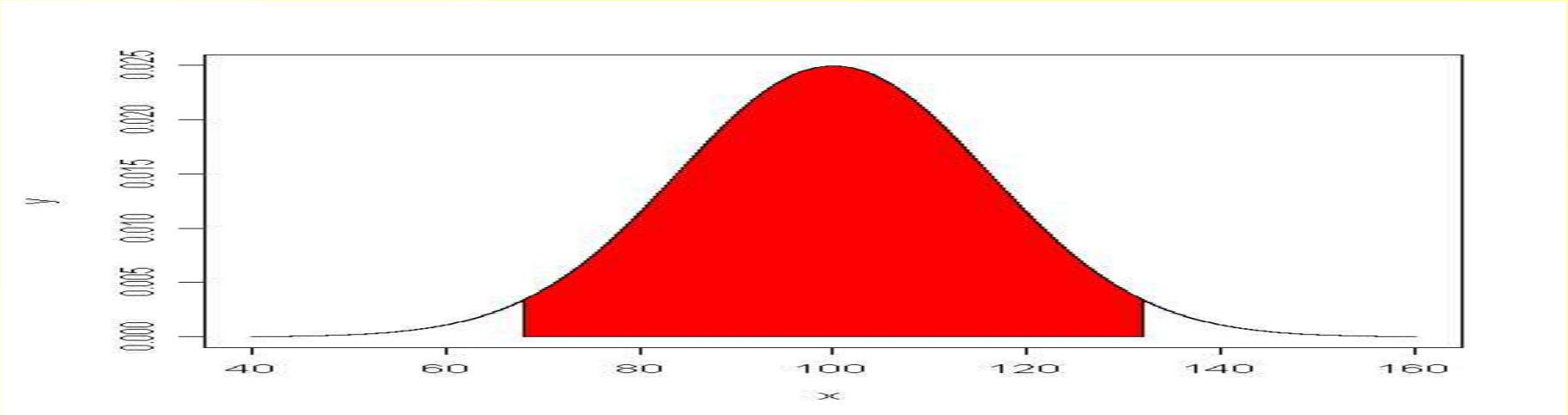
- Normal distributions have a nice property that, knowing the mean (μ) and standard deviation (σ), we can tell how much data will fall in any region.
- In other words: The complete distribution is determined by the two parameters.
- Examples – the normal distribution is symmetric, so 50% of the data is smaller than μ and 50% is larger than μ.

Verifying the Empirical Rule

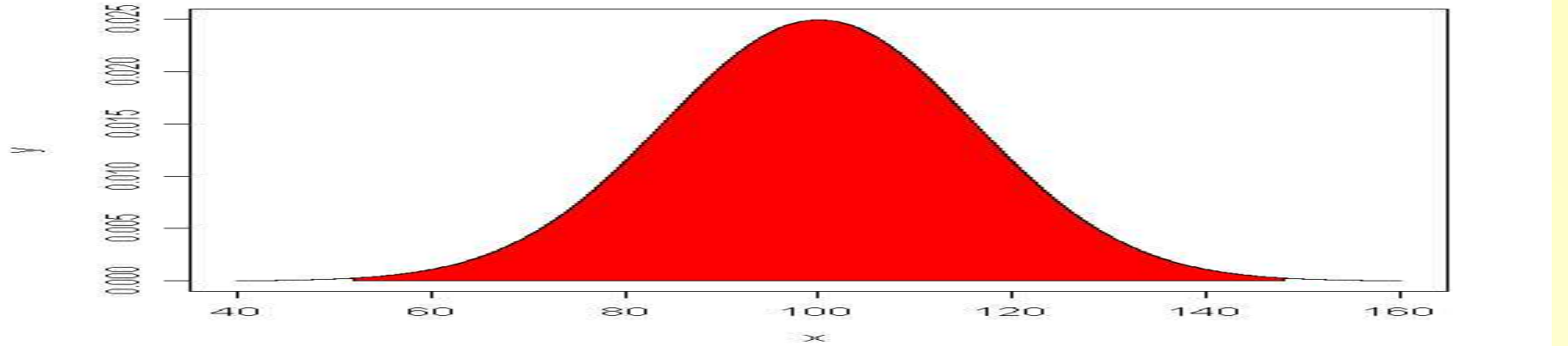
- It is always true that about 68% of the data appears within 1 standard deviation of the mean (so about 68% of the data appears in the region $\mu \pm \sigma$)
- Normal Density Curve Applet



95% within 2 standard deviations



99.7% of the data (almost all the data) within 3 standard deviations of the mean



- In quality control applications, one often is interested in “6-sigma”.
- 6 standard deviations include 99.9999998% of the data.

Normal Distribution: Example (female height)

- Assume that adult female height has a normal distribution with mean $\mu=165$ cm and standard deviation $\sigma=9$ cm
- With probability 0.68, a randomly selected adult female has height between

$$\mu - \sigma = 156 \text{ cm and } \mu + \sigma = 174 \text{ cm}$$

- With probability 0.95, a randomly selected adult female has height between

$$\mu - 2\sigma = 147 \text{ cm and } \mu + 2\sigma = 183 \text{ cm}$$

- Only with probability 1-0.997=0.003, a randomly selected adult female has height below

$$\mu - 3\sigma = 138 \text{ cm or above } \mu + 3\sigma = 192 \text{ cm}$$

Normal Distribution

- So far, we have looked at the probabilities within one, two, or three standard deviations from the mean

  $(\mu + \sigma, \mu + 2\sigma, \mu + 3\sigma)$
- How much probability is concentrated within 1.43 standard deviations of the mean?
- More general, how much probability is concentrated within *z* standard deviations of the mean?

Normal Distribution Calculators

- Many statistics textbooks contain tables of the normal distribution probabilities
- Online tools are easier to use, and more precise
- Standard Normal Calculator "Surfstat"
- Standard Normal Calculator "Stat Trek"

- Example, for z=1.43:

  The probability within 1.43 standard deviations of the mean is _____.

  The probability outside 1.43 standard deviations of the mean is _____.

Working backwards

- We can also use the online calculator to find *z*-values for given probabilities
- Find the *z*-value corresponding to a right-hand tail probability of 0.025
- Answer: Probability 0.025 lies above

  $\mu +$ _____ σ
- Find the *z*-value for a right-hand tail probability of 0.1, 0.05, 0.01

Finding *z*-Values for Percentiles

- For a normal distribution, how many standard deviations from the mean is the 90^{th} percentile?
- Or: What is the value of z such that 0.90 probability is less than $\mu + z\,\sigma$?
- Answer: The 90^{th} percentile of a normal distribution is _____ standard deviations above the mean

Application

- SAT scores are approximately normally distributed with mean 500 and standard deviation 100
- The 90^{th} percentile of the SAT scores is 1.28 standard deviations above the mean
- $\mu + 1.28\,\sigma = 500 + 1.28 \cdot 100 = 628$
- Find the 99^{th} and the 5^{th} percentile of SAT scores

Online Tools

http://bcs.whfreeman.com/scc/content/cat_040/spt/normalcurve/normalcurve.html
http://stat.utilities.googlepages.com/tables.htm
http://stattrek.com/Tables/Normal.aspx

- Use these to
  - verify graphically the empirical rule,
  - find probabilities,
  - find percentiles
  - calculate *z*-values for one- and two-tailed probabilities

Example

- In baseball, batting average is the number of hits divided by the number of at-bats.
- Recent batting averages of almost 1000 Major League Baseball players could be described by a normal distribution with mean 0.270 and standard deviation 0.008.
- What percent of the players have a batting average of 0.28 and greater?
- What percentage have a batting average of below 0.25?
- If there are 30 players on a roster, how many would you expect to have a batting average of above 0.28 (below 0.25)

Another Example

- Assume that cholesterol levels of men in the US have an approximately normal distribution with mean 215 (mg/dl) and standard deviation 25 (mg/dl).
- What is the probability that the cholesterol level of a randomly selected man is less than 180?
- What is the probability that it is between 190 and 220?

Quartiles of Normal Distributions

- Median: z=0

  (0 standard deviations above the mean)
- Upper Quartile: z = 0.67

  (0.67 standard deviations above the mean)
- Lower Quartile: z= – 0.67

  (0.67 standard deviations below the mean)
- Find the lower and upper quartile of cholesterol levels for men in the US

z-Scores

- The *z*-score for a value x of a random variable is the number of standard deviations that x is above μ
- If x is below μ, then the *z*-score is negative
- The *z*-score is used to compare values from different normal distributions

Calculating *z*-Scores

- You need to know x, μ, and σ
  to calculate *z*

$$z = \frac{x - \mu}{\sigma}$$

Tail Probabilities

- SAT Scores: Mean=500,
  Standard Deviation =100
- The SAT score 700 has a *z*-score of *z*=2
- The probability that a score is ***beyond*** 700 is the tail probability of *z*=2
- Online tool….
- 2.28% of the SAT scores are ***beyond*** 700 (***above*** 700)

Tail Probabilities

- SAT score 450 has a *z*-score of *z*=-0.5
- The probability that a score is ***beyond*** 450 is the tail probability of *z*=-0.5
- Online tool….
- 30.85% of the SAT scores are ***beyond*** 450 (***below*** 450)

z-Scores

- The *z*-score is used to compare values from different normal distributions
- SAT: μ=500, σ=100
- ACT: μ=21, σ=6
- What is better, 650 in the SAT or 28 in the ACT?

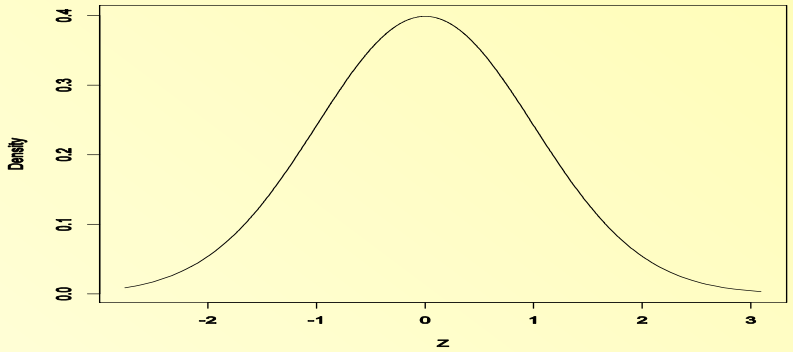
$$z_{SAT} = \frac{x - \mu}{\sigma} = \frac{650 - 500}{100} = 1.5$$

$$z_{ACT} = \frac{x - \mu}{\sigma} = \frac{28 - 21}{6} = 1.17$$

Corresponding tail probabilities?
How many percent have better SAT or ACT scores?

Standard Normal Distribution

- The standard normal distribution is the normal distribution with mean $\mu=0$ and standard deviation $\sigma=1$



Standard Normal Distribution

- When values from an arbitrary normal distribution are converted to *z*-scores, then they have a standard normal distribution
- The conversion is done by subtracting the mean μ, and then dividing by the standard deviation σ

$$z = \frac{x - \mu}{\sigma}$$

Example

- 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.
- Find the probability that PDI is between 97 and 103.
- Find the *z*-score for a PDI value of 90. Would you be surprised to observe a value of 90?
- Suppose we convert all the PDI observations to *z*-scores; that is, for each infant, subtract 100 from the value of PDI and divide by 15. Then, what is the distribution of the *z*-scores called? What are the mean and standard deviation of these *z*-scores?

Typical Questions

- One of the following three is given, and you are supposed to calculate one of the remaining
  1. Probability or percentage (right-hand, left-hand, two-sided, middle)
  2. z-score
  3. Observation *x,* original score
- In converting between 1 and 2, you need one of the online tools.
- In transforming between 2 and 3, you need mean and standard deviation and one of the following formulas

$$z = \frac{x - \mu}{\sigma} \qquad x = \mu + z\sigma$$

Note: Most of the time, mu and sigma are provided. If not, things can be a bit more tricky.

Online Tools

http://bcs.whfreeman.com/scc/content/cat_040/spt/normalcurve/normalcurve.html
http://stat.utilities.googlepages.com/tables.htm
http://stattrek.com/Tables/Normal.aspx

- Use these to
  - verify graphically the empirical rule,
  - find probabilities,
  - find percentiles
  - calculate *z*-values for one- and two-tailed probabilities

More Z-Score Examples

- IQ Scores: $\mu=100$ and $\sigma=15$
- An observation $X=125$ is 25 points above the mean, which corresponds to $25/15 = 1.67$ standard deviations above the mean.
- In general, a Z-score for an observation X is $\mathbf{Z=(X-\mu)/\sigma}$
- Observations above the mean get positive Z-scores, observations below the mean get negative Z-scores.

From Percentiles to Z-Scores

- What is the 80^{th} percentile of IQ Scores?
- In other words, for what IQ do 80% of the people fall below it?
- The first step is to find the Z-score corresponding to 80%
- That Z-score is Z=0.84

From Z-scores to IQ Scores

- The Z-score corresponding to 80% is 0.84
- The Z-score formulas are

$Z=(X-\mu)/\sigma$ and $X = \mu + \sigma Z$

- $Z=0.84$, $\mu=100$, and $\sigma=15$, so
- $X = 100+(0.84)(15) = 112.6$
- So 80% of people have an IQ score below 112.6

Middle Percentages

- What about the middle 50% of IQ scores?
- What percentiles does this correspond to? To get the middle 50%, we need to stretch from the 25^{th} percentile to the 75^{th} percentile.
- The 25^{th} percentile corresponds to a Z-score of _____

  the 75^{th} percentile corresponds to a Z-score of _____
- These correspond to IQ's of _____ and _____
- What about the middle 95% of values (removing 2.5% from each tail)
- Finally, the middle 99% of IQ scores?