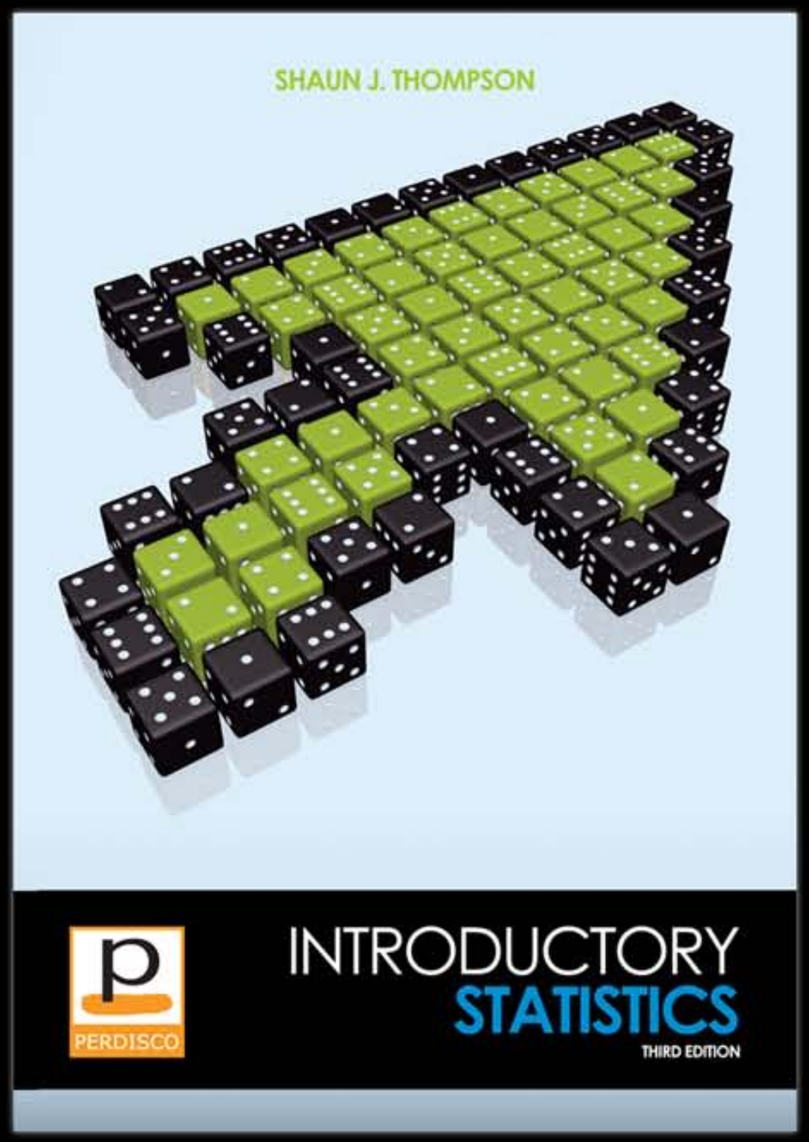


These lecture slides are designed to accompany:

Introductory Statistics, Third Edition

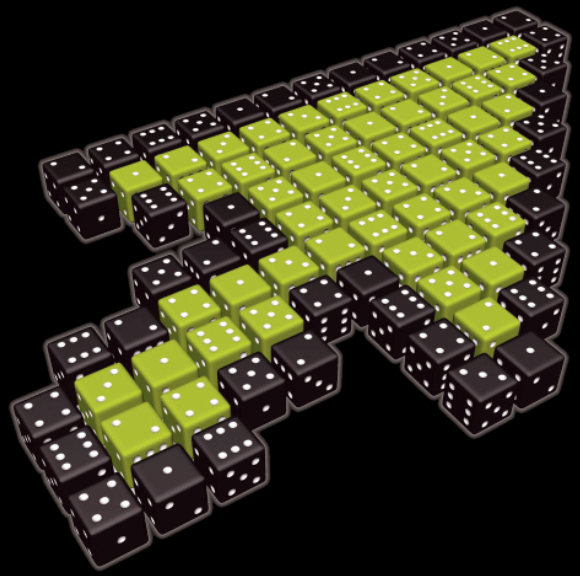
Other features include:

- Chapter summary videos
- MP3 audio podcasts
- VirtualTutor e-learning
- AntiCheat and AutoGrade homework
- Detailed instructor resources

To find out more, visit: perdisco.com/stats

Chapter 5

Probability distributions



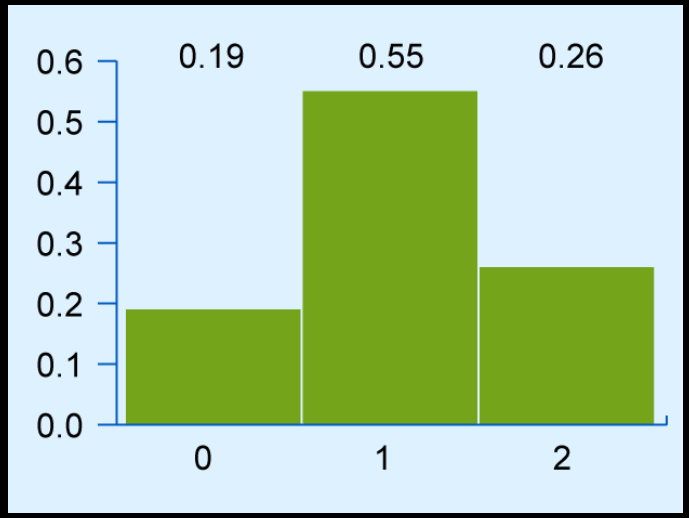
Data versus variables

- When we collect data, we are observing a variable assume different values
- When we present and measure that data, we are describing the values that the variable assumed
- So descriptive statistics is about what a variable did
- We would like to describe what the variable can do
- This will involve probability
- In fact, it will involve **probability distributions**!

Example – data distribution

- Flip a coin twice and count the number of heads
- This gives a variable with 3 different values: 0, 1, 2
- Repeat the process 100 times → 100 data values
- We can describe the distribution of this data:

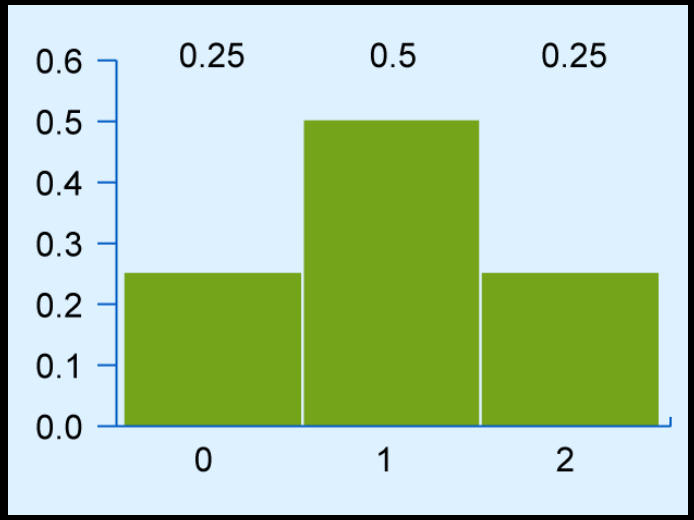
Number of heads	Relative frequency
0	0.19
1	0.55
2	0.26



Example – probability distribution

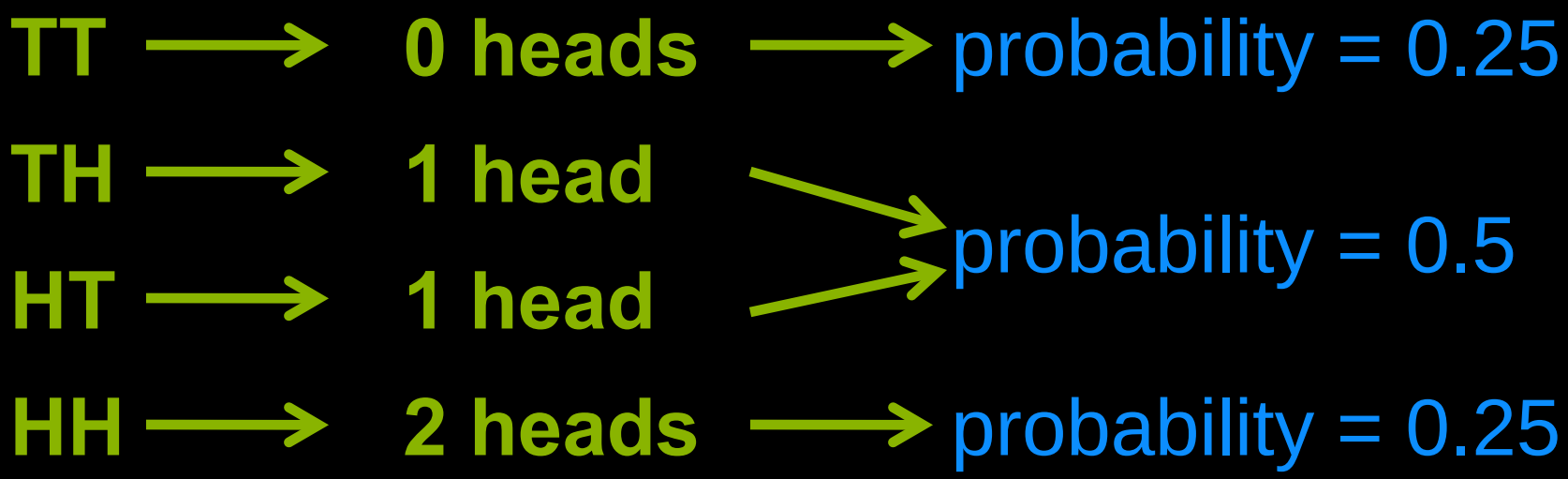
- Instead, of repeating 100 times, just think about probabilities of different values
- We can describe their distribution too:

Number of heads	Probability
0	0.25
1	0.50
2	0.25



Example – probability distribution (cont'd)

- Why are these the probabilities?
- Because two outcomes lead to 1 head occurring:



- This is why these three values of the variable get assigned these probabilities

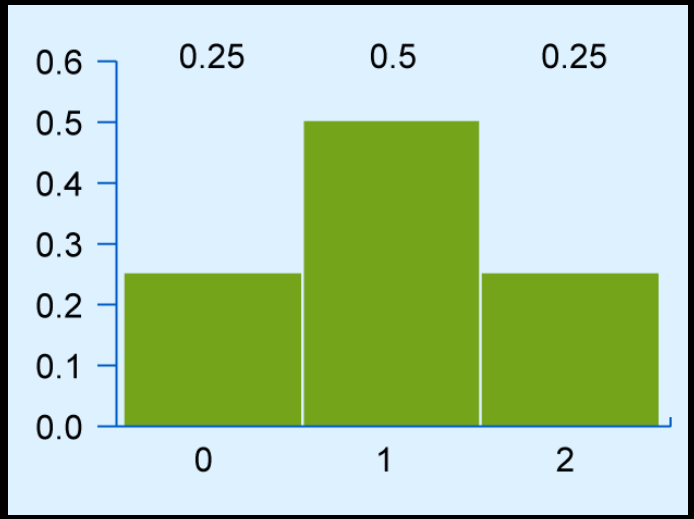
Random variables

- When interested in the probabilities of the values in a variable, instead of observed data values, we call it a random variable
- More precisely: a **random variable** is a variable **X** that can assume a numerical value for every outcome in a probability sample space
- **Discrete** random variable
  - assumes values from a discrete list
- **Continuous** random variable
  - assumes values from a continuous spectrum

Probability distribution

- Assigns a probability p(x) to each value x of a discrete random variable X
- Example – coin flip experiment:
  - p(0) = 0.25, p(1) = 0.5, p(2) = 0.25

x	p(x)
0	0.25
1	0.50
2	0.25



Measuring random variables

- Just like we can measure sets of observed data, we can measure the variables they come from
- In particular, we can measure the center and spread in a variable X
- The center is measured by a number known as the **expected value** of X
- The spread is measured by the **variance** and **standard deviation** of X
- Remember: these measure the underlying variable!

Expected value

- For a discrete random variable X
  - values $x_1, \ldots, x_n$ and probabilities $p(x_1), \ldots, p(x_n)$
- **Expected value** is:

$$E(X) = \sum_{i=1}^{n} x_i p(x_i)$$

- Sometimes referred to as the **mean** of X, denoted μ

Variance and standard deviation

- For a discrete random variable X
  - values $x_1, \ldots, x_n$ and probabilities $p(x_1), \ldots, p(x_n)$
- **Variance** is:

$$VAR(X) = \sum_{i=1}^{n} (x_i - E(X))^2 p(x_i)$$

- **Standard deviation** is:

$$SD(X) = \sqrt{VAR(X)}$$

Binomial distribution

- Suppose you want to flip a coin 10 times and count the number of heads
  - What is the probability of getting exactly 5 heads?
  - What is the probability of getting between 4 and 6 heads (inclusive)?
  - For any x between 0 and 10, what is the probability of getting exactly x heads?

- The **binomial distribution** can answer these!

The scenario for the binomial distribution

- The binomial distribution applies if:
  1. You repeat a process (trial) that can result in two mutually exclusive and collectively exhaustive outcomes (success and failure)
  2. The trial is repeated a fixed number of times, n
  3. Probability of success each time is fixed, p
  4. Trials are independent
- The binomial distribution is a formula p(x) for the probability of getting x successes in these trials
- It can be used for any x between 0 and n

Formula for binomial distribution

- For n trials and probability of success p, the formula for the probability of x successes is:

$$p(x) = {}^{n}C_{x}\, p^{x} (1-p)^{n-x}$$

- This is what is known as the **binomial distribution**
- Note that:

$${}^{n}C_{x} = \frac{n!}{x!(n-x)!}$$

Example

- In the coin flip example, the trial is flipping the coin
- Success = head, Failure = tail
- Number of trials is $n = 10$
- Probability of success is $p = 0.5$
- Formula for probability of x heads out of 10 is:

$$p(x) = {}^{10}C_x \, 0.5^x 0.5^{10-x}$$

Example continued

- Probability of exactly 5 heads:

$$p(5) = {}^{10}C_5 \; 0.5^5 \; 0.5^{10-5}$$

$$= \frac{10!}{5! \times 5!} \times 0.03125 \times 0.03125$$

$$\approx 0.2461$$

$$= 24.61\%$$

Measures of binomial distribution

- If X follows the binomial distribution with n trials and probability p of success, then:

$$E(X) = np$$

$$VAR(X) = np(1-p)$$

$$SD(X) = \sqrt{np(1-p)}$$

- So the measures of the binomial distribution will depend on n and p

Continuous random variables

- Everything so far has been about discrete variables
- What about continuous variables?
- Example
  - the time taken to finish a task is a continuous variable
- The way we assign probabilities is different
- For example, every individual outcome is assigned a probability of zero!

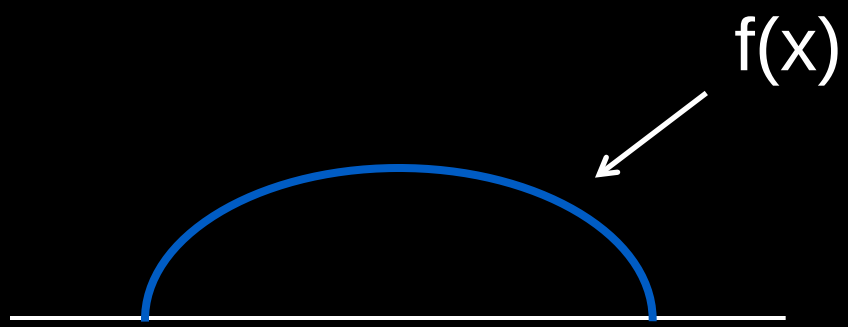
Assigning probabilities

- If X is a continuous random variable, then its values exist along a continuous spectrum
- We don't assign probabilities to individual values
- Instead, we assign probabilities to regions of the spectrum
- This is done with a function called the **probability density function, f(x)**

Probability density function

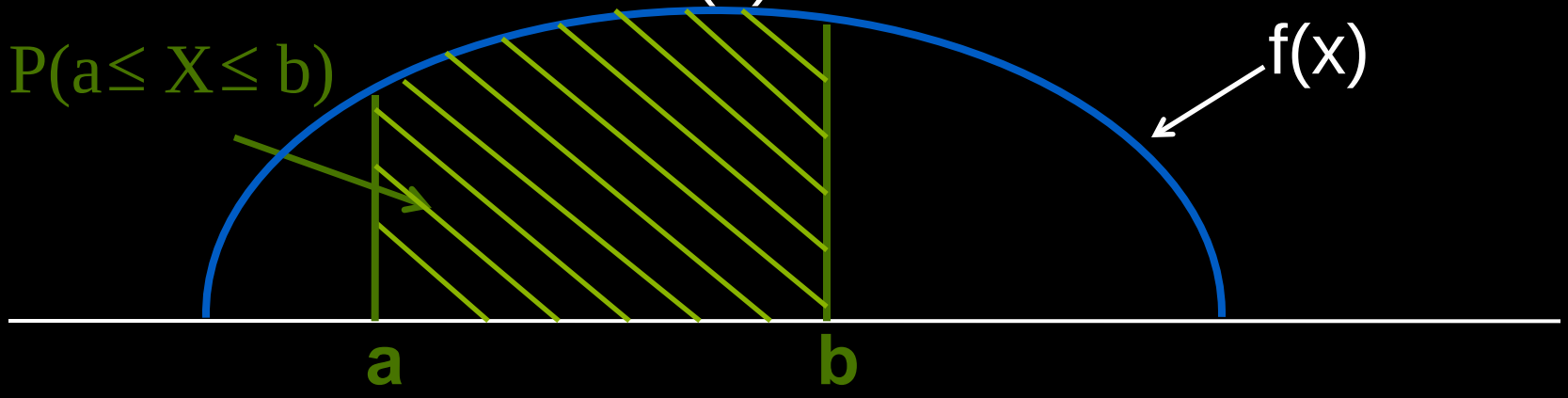
- A probability density function is a function f(x) that assigns numbers to the values of X such that:

1. The numbers are always positive, i.e., the curve of f(x) is always above the x-axis
2. The total area underneath the curve f(x) is always 1



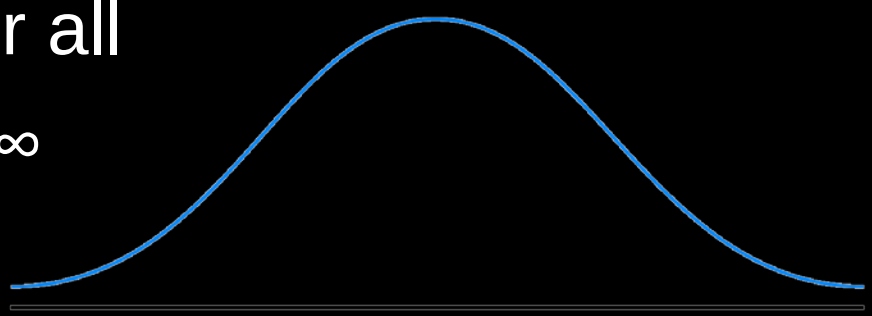
Using a probability density function

- The numbers assigned by a probability density function are not probabilities!
- However, they are related to probabilities in X
- For any two values a and b of X, the probability that X will assume a value between a and b is the area underneath the curve f(x) between a and b



Normal distribution

- By far the most important and commonly-used distribution in statistics!
- The **normal distribution** occurs everywhere in nature, science, business, psychology, etc.
- It is continuous and has a smooth bell-shaped curve that you may have seen before
- The curve is defined over all values between $-\infty$ and ∞

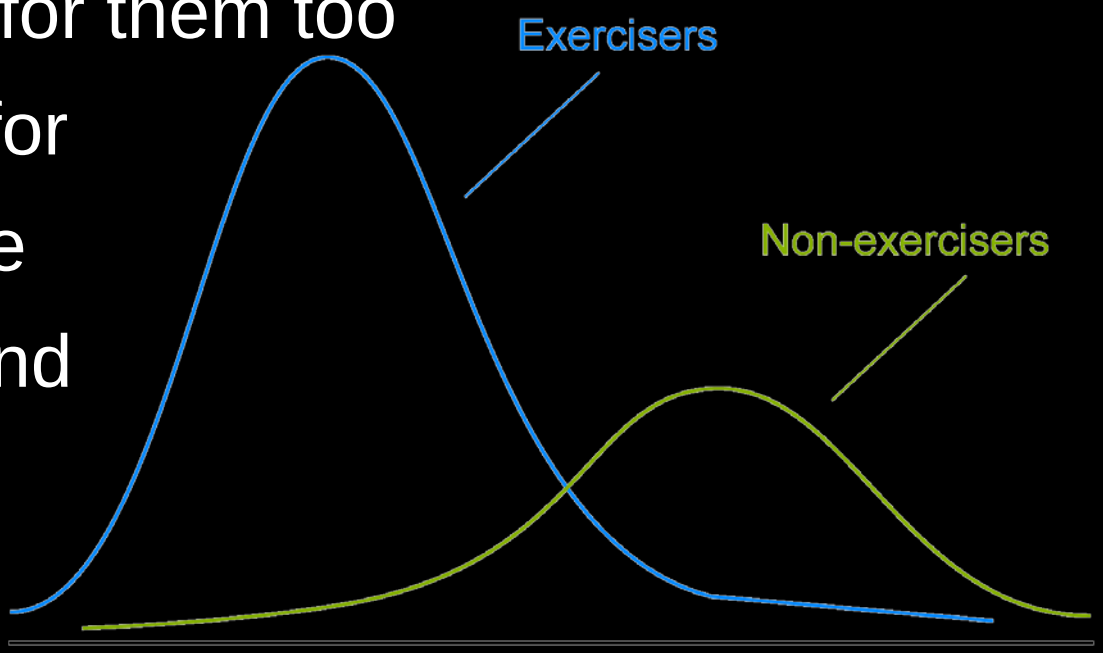


Parameters of the distribution

- There is a whole family of different normal distributions
- Each distribution is specified by two parameters:
  - the mean, μ
  - the standard deviation, σ
- These determine the exact shape and position of the bell curve
- Larger μ → peak of bell curve further to the right
- Larger σ → bell curve is ‘shorter and fatter’

Example

- Heart rate in exercisers versus non-exercisers
- Mean heart rate is lower for exercisers, and heart rate is more consistent for exercisers, so standard deviation is lower for them too
- So the bell curve for exercisers is to the left, and is taller and thinner



Probability density function

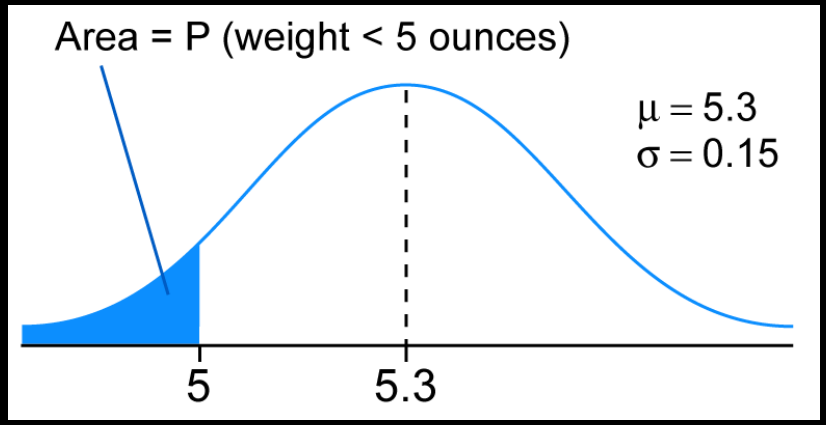
- Here is the probability density function for the normal distribution X with mean μ and standard deviation σ:

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

- Quite nasty!
- Don't worry – you won't have to use it!

Probabilities in the normal distribution

- So how do we answer probability questions?
- Example: A chocolate factory makes chocolates with mean weight 5.3 ounces, standard deviation 0.15 ounces. Weights follow a normal distribution.
- If factory discards all chocolates below 5 ounces, what proportion does it discard?



Standard normal distribution

- We can answer questions like this by using the **standard normal distribution**, Z
- Z is the normal distribution with:
  - mean $\mu = 0$
  - standard deviation $\sigma = 1$
- Probability density function for Z:

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}$$

Standard normal table

- That is still fairly nasty!
- But someone has done the hard work for us

- The **standard normal table** is used to find probabilities from the standard normal distribution Z

- Most directly, for any value z, the table can be used to find $P(Z < z)$, the probability that Z will assume a value less than z

Finding probabilities in Z

- That probability that Z will assume a value less than 0.71 is 76.11%

z	0.00	0.01	0.02	0.03	0.04
0.6	0.7257	0.7291	0.7324	0.7357	0.7389
0.7	0.7580	**0.7611**	0.7642	0.7673	0.7704
0.8	0.7881	0.7910	0.7939	0.7967	0.7995

- We can also use this table for negative values z
- Z is symmetric, so the probability that Z will assume a value less than −0.71 is given by:

$$P(Z<-0.71)=1-P(Z<0.71)$$

$$=23.89\%$$

- Software can also be used to calculate probabilities

Transformation formula

- But how does this help the chocolates question?
- Any normal distribution X can be transformed into the standard normal distribution Z:

$$Z = \frac{X - \mu}{\sigma}$$

- This formula can also be used to transform values of any X into values of Z

Example

- In the chocolate example X had parameters $\mu = 5.3$ and $\sigma = 0.15$, so the value 5 is equivalent to:

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

$$= \frac{5 - 5.3}{0.15}$$

$$= -2$$

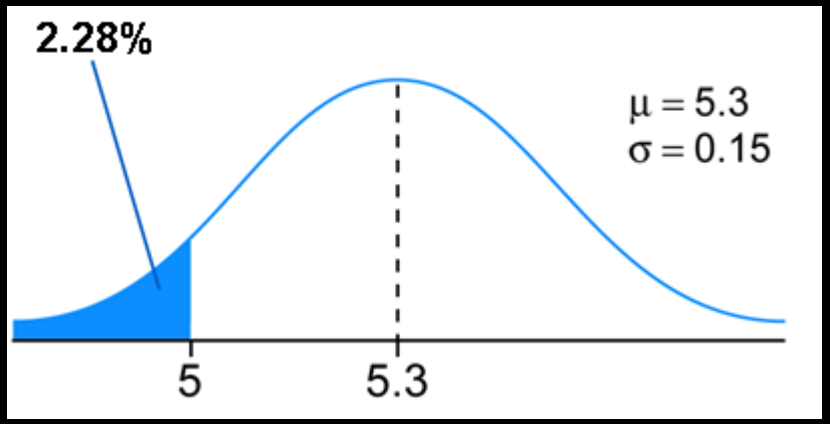
- The value -2 is known as the **z-score** of 5

Example continued

- The idea is that the value 5 in the chocolate distribution is the ‘same as’ the value -2 in the standard normal distribution
- We wanted to know what proportion of chocolates were less than 5 ounces
- This becomes equivalent to asking: What proportion of Z falls below -2?
- The standard normal table (or software) can be used to find that 2.28% of Z falls below -2

Example continued

- This is equivalent to saying that 2.28% of X falls below 5
- Therefore 2.28% of the chocolates will have weights less than 5 ounces!
- This is the proportion of chocolates that will be discarded



Areas within the normal distribution

- Some useful facts about the normal distribution:
  - Around 90% of the values will fall within 1.645 standard deviations of the mean
  - Around 95% of the values will fall within 1.960 standard deviations of the mean
  - Around 99% of the values will fall within 2.576 standard deviations of the mean
- These facts come in very handy in statistical inference later on

Normal approximation to binomial

- Recall that the binomial distribution applies to a discrete random variable X
- Distribution is specified by two parameters, n and p
- For large number of trials, n, questions about the distribution can become hard
- Example: Suppose p = 30% of people donate blood, and a sample of n = 1,000 people are asked if they donate blood.
- What is the probability that at least 320 say yes?

Normal approximation to binomial cont'd

- This could technically be answered by plugging every value from 320 up to 1000 into the binomial distribution, and adding the answers up
- This is not very fun to do!
- Alternative: assume that X is actually normal with the same mean and standard deviation as this binomial distribution
  - $\mu = np = 1{,}000 \times 0.3 = 300$
  - $\sigma = \sqrt{(np(1 - p))} = \sqrt{(1{,}000 \times 0.3 \times 0.7)} = 14.4914$

Normal approximation to binomial cont'd

- Now the question becomes:

  In the normal distribution X with $\mu = 300$ and $\sigma = 14.4914$, what proportion of values fall above 320?

- Actually, we want to include the value of 320 in the original question (but not 319), so we should ask:

  What proportion of X falls above 319.5?

  – This is known as the correction for continuity

- Now the answer is 8.92% - this is the chance that at least 320 people in the sample do donate blood
  - Note: without using the normal approximation, the correct answer is 8.98% - so our approximate answer is quite close!