Lecture notes 4: Probability and the normal distribution

Highlights:

• Random events
• Interpreting probability
• Standardization and z-scores
• Chebyshev’s rule
• The normal distribution
• The empirical rule
• Calculating probabilities based on the normal distribution
• Percentiles
• Calculating percentiles based on normal probabilities
Random Events

• “Random” is a difficult word to define. In statistics, we use it to refer to events that are unpredictable.

• A random event is something that may or may not occur, and that which we can assign a probability to.

• We can think of a random event as a possible value that a random variable takes on. For instance, a random variable “x” may be the outcome of a roll of a die, and a random event might be x = 6.

• Generally we will denote a random variable as “x”. If a random variable follows a known distribution, we will refer to that distribution (e.g. “z”, “t”, “F”)
Examples of random events

- You flip a coin and it comes up heads
- It snows tomorrow
- The Broncos win the Super Bowl next year
- The Dow Jones finishes the week up more than 50 points

The complement of an event is its non-occurrence, or its opposite. For example:

- You flip a coin and it comes up tails
- It does not snow tomorrow
- The Broncos fail to win the Super Bowl next year
- The Dow Jones finishes the week either down, or up less than 50 points
Probability

• Probability is a way of quantifying the likelihood (i.e. chance) that some random event occurs.

• Probabilities are often related as percentages, but formally they should be given as proportions.

• For example, if there is a 50% chance of something happening, then its probability is 0.5.

• A probability MUST be a number between 0 and 1. Think of 0 as “improbable” and 1 as “absolutely certain”.
“Relative frequency”

- Relative frequency expresses how often an event occurs as a proportion of how often it potentially could have occurred.
- We saw this when we looked at histograms, where relative frequency was expressed a proportion of a sample. But relative frequency doesn’t have to refer to a sample.
- The “relative frequency” interpretation of probability states that the probability of a specific outcome is the proportion of times the outcome would occur over the long run (i.e. if we kept repeating the random process indefinitely).
“Relative frequency”

- For example, if we say that the probability of getting “heads” on a coin flip is 0.5, what we mean is that if we were to flip the coin 1000 times, we would expect to get heads 500 times.

- If probability = 0.5, then no matter how many flips we make, we always expect half to come up heads.

- Note that this version of relative frequency does not need to involve a sample size. It does not need to involve a denominator. It is just a proportion that can be applied to any sample size.
Probability of Male versus Female Births

- Another example: long-run relative frequency of males born in the United States is about .512.

  *From the Information Please Almanac (1991, p. 815).*

- Here is some data taken from hospital records:

```
<table>
<thead>
<tr>
<th>Weeks of Watching</th>
<th>Total Births</th>
<th>Total Boys</th>
<th>Proportion of Boys</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>30</td>
<td>19</td>
<td>.633</td>
</tr>
<tr>
<td>4</td>
<td>116</td>
<td>68</td>
<td>.586</td>
</tr>
<tr>
<td>13</td>
<td>317</td>
<td>172</td>
<td>.543</td>
</tr>
<tr>
<td>26</td>
<td>623</td>
<td>383</td>
<td>.615</td>
</tr>
<tr>
<td>39</td>
<td>919</td>
<td>483</td>
<td>.526</td>
</tr>
<tr>
<td>52</td>
<td>1237</td>
<td>639</td>
<td>.517</td>
</tr>
</tbody>
</table>
```
Probability of Asparagus Hypersensitivity

• According to a 2012 study, 1 out of every 20 people have a gene that makes them hypersensitive to the smell of asparagus.

• This number is based on data collected over the long run. So the probability that a randomly selected individual in 2012 would have this hypersensitivity to asparagus is $1/20$, or 0.05
Expressing relative frequency

Here are a few different ways of expressing this probability / relative frequency:

- The proportion of people with asparagus hypersensitivity is 1/20 or .05

- 5% of people have asparagus hypersensitivity.

- The probability that a randomly selected person has asparagus hypersensitivity is .05

- If we sampled 1000 people at random, we expect that 50 of them would be hypersensitive to asparagus.
Probability notation

• P(X) is the probability that event X occurs.

• 1 - P(X) is the probability that event X does not occur.

• In other words, the probability of the complement of X is 1 - P(X).

• For example, if X is the event that it rains tomorrow, and the probability of rain tomorrow is 0.3, then the probability it does not rain tomorrow is 1 - P(X) = 1 - 0.3 = 0.7
Standardization and z-scores

• In class example #2, we introduced the concept of standardization.

• When values are standardized, they are put into some kind of common unit.

• In statistics, we commonly standardize an observed value of a variable by finding the distance from its mean in terms of standard deviations.

• We call a value that has been standardized in this manner a z-score.
Standardization and z-scores

• Formally, we use this formula to standardize a value into a z-score:

\[ z = \frac{x - \mu}{\sigma} \]

• \( x \) is the value to be standardized

• \( \mu \) ("mu") is the population mean

• \( \sigma \) ("sigma") is the population standard deviation

• Note that \( \mu \) and \( \sigma \) are population parameters. Parameters are typically denoted using Greek letters.
Standardization and z-scores

- If the population mean and standard deviation are unknown, we can use the sample mean and standard deviation:

\[ Z = \frac{x - \bar{x}}{s} \]

- \( x \) is the value to be standardized
- \( \bar{x} \) is the sample mean
- \( s \) is the sample standard deviation

Note that \( \bar{x} \) and \( s \) are sample statistics. Statistics are typically denoted using English letters.
Z-score example

• Recall the snake undulation rate example from lecture notes 3.

• In this example, we calculated a sample mean of 1.375 Hz, with a corresponding standard deviation of .324

• The largest observation in this data set was 2 Hz. Let’s covert this to a z-score and determine how much 2 deviates from the mean undulation rate, in terms of standard deviations:
Interpretation of Z-scores

- The sign of the z-score indicates whether the data point lies above or below the mean.

- A positive z-score means that the observation is above the mean and a negative z-score means that the observation is less than the mean.

- The magnitude of the z-score indicates how far away the data point is from the mean in terms of standard deviations.

- We sometimes say that z-scores are “unitless”, but really their units are # of standard deviations from the mean.
Chebyshev’s Rule

• To get an idea of how many standard deviations from the mean is “a lot”, we can use **Chebyshev’s Rule** (a.k.a Chebyshev’s Inequality or Chebyshev’s Theorem)

• Chebyshev’s rule states that at least

\[
\left(1 - \frac{1}{k^2}\right) \cdot 100\%
\]

of the values in a distribution must lie within \( k \) standard deviations of its mean.
Chebyshev’s Rule example

• So, if we want to answer the question: “at least what percentage of a dataset must be within 2 standard deviations of the mean?”, we can simply plug $k=2$ into Chebyshev’s formula:

• Likewise with 3 standard deviations:
Chebyshev’s Rule example

• And 4 standard deviations:

• Seen in this context, one standard deviation is actually a pretty big quantity.
The Normal Distribution

- Chebyshev’s rule applies to ANY distribution.

- If we know the shape of a distribution, we can make an even stronger statement about how much data falls within a certain number of standard deviations from the mean.

- The most common distribution we see in statistics is the normal distribution. This is also known as the Gaussian distribution or the “bell curve”.

- A variable that follows a normal distribution is said to be normally distributed.
Examples of normal distributions

![Image of normal distributions and SAT Scores in 2010 graph]
The Normal Distribution

• The normal distribution is bell shaped, and it is defined by its mean ($\mu$) and its variance ($\sigma^2$) or standard deviation ($\sigma$).

• We will spend a lot of time talking about the properties of the normal distribution, and how we use it to compute probabilities.

• One useful property of normal distribution is given by the **empirical rule**, which is just like Chebyshev’s rule, but tailored specifically to normal distributions.
The Normal Distribution

• Below is a special case of the normal distribution, called the **standard normal distribution**. It is defined to have a mean of $\mu=0$ and a standard deviation of $\sigma=1$. It’s horizontal axis is $z$. 
The empirical rule

- The empirical rule tells us what percentage of the values of a normally distributed variable fall within 1, 2, and 3 standard deviations of the mean.

- The empirical rule is analogous to Chebyshev’s rule, but it only applies to normal distributions.

- Because it assumes that a variable is normally distributed, it can be used to make stronger statements about how much data must be within a certain number of standard deviations of the mean.
The empirical rule

- By the empirical rule, about 68% of the values in a normal distribution will be within one standard deviation of the mean:
By the empirical rule, about 95% of the values in a normal distribution will be within two standard deviations of the mean:
The empirical rule

- By the empirical rule, about 99.7% of the values in a normal distribution will be within three standard deviations of the mean:

\[ \mu - 3\sigma \] to \[ \mu + 3\sigma \]

The shaded area is 99.7% of the total
The empirical rule example

- Suppose we know that the mean height of adult women in the U.S. is normally distributed, with a mean of 64 inches, and a corresponding standard deviation of 2.75 inches.

- Using the empirical rule, we can say that about 99.7% of all U.S. adult women are between what two heights?
Normal probability

• The normal distribution is a type of **probability distribution**. A probability distribution shows us the values that a variable takes on, and how likely it is that it takes those values on.

• We define the area under a probability distribution to equal 1. Then, **we can use this area to represent probabilities**.

• In the case of a normally distributed variable, we can convert it to a z-score so that it will follow the standard normal distribution, which is also called the **z distribution**.
Normal probability side note

• A “z-score” is, in general, any standardized value.

• The “z-distribution” is a normal distribution whose values have been standardized.

• It is important to note that standardizing the values of a variable does not make that variable normal!

• The z-distribution should only be used to calculate probabilities when the variable in question is known to be normally distributed.
Normal probability example

• Suppose we have a variable that we know is normally distributed, such as height. The mean and standard deviation of U.S. adult male heights are 70” (5’10”) and 3”, respectively. We can use this information to answer questions, such as:

• What is the probability that a man will be less than 64 inches (5’4”) tall?

• What is the probability that a man will be between 61 inches (5’1”) and 76 inches (6’4”) tall?

• What height separates the bottom 10% from the top 90% of male heights?
Normal probability example

• To answer these questions, we convert whatever height we are interested in to a z-score.

• We can then find the area under the standard normal distribution that corresponds to the probability we are interested in.

• So, to find the probability that a man will be less than 64 inches tall, we can plug 64 into our z formula:
Normal probability example

• Now we can label this z-value on the standard normal distribution, and shade the area corresponding the probability we are interested in:
Normal probability example

- This shaded area represents the probability that a man will be less than 64 inches tall. Calculating it by hand is extremely complicated, but thankfully we can do it with our calculator.

- In a TI-83 or TI-84, go to DIST (using 2nd VARS), then select normalcdf().

- This function will give you the area under the standard normal distribution between any two z-values. The syntax is:

  - normalcdf( left z-value, right z-value )
Normal probability example

• Since in this case we want the area to the left of $z = -2$, we will need to pick a “far” left endpoint that is so far out that there is practically no area under the curve above it. Let’s use $z = -100$.

• Enter `normalcdf(-100,-2)`, and your calculator will tell you how much area there is under the standard normal distribution between $z = -100$ and $z = -2$.

• This area is the probability that a randomly selected American male will be less than 64 inches tall.
Normal probability example 2

• Let’s answer the next question using this same method:

  What is the probability that a man will be between 61 inches (5’1”) and 76 inches (6’4”) tall?

• Again, we first use the z formula to standardize these heights:
Normal probability example 2

- Now label these z-values on the normal curve, and this time shade the area between them, since that represents the probability that we are interested in.
Normal probability example 2

• Now use the normalcdf() function to find the probability represented by the shaded area. In this case, the left and right endpoints are the two z-scores that we calculated:
Normal probability example 3

• Suppose we want to know what proportion of U.S. adult males are taller than 6 feet (72 inches)

• Convert 72 inches to a z-score, label it on the normal curve, and shade the corresponding area:
Now, since we want the area to the right of $z = 0.67$, we should use 0.67 as our left endpoint and some very large value as our right endpoint. Let’s use $z = 100$.

\[
\text{normalcdf}(0.67,100) =
\]

This is the proportion of U.S. male adults who are over 6 feet tall.

(assuming of course that our mean and standard deviation are correct, and that height is normally distributed)
Percentiles

• We can also do this process in reverse: instead of taking an x value and computing a probability or proportion from it, we can take a probability or proportion and find the x value to which it corresponds.

• For instance, we may want to know what height separates the shortest 10% of men from the tallest 90% of men. In other words, we want to know the 10\textsuperscript{th} percentile of male heights.

• In general, a percentile is the value of a variable for which the given percentage of values fall below it.
Percentiles

• For example, the median is the 50\textsuperscript{th} percentile, because 50\% of the data falls below the median.

• Similarly, Q1 is the 25\textsuperscript{th} percentile, because 25\% of the data falls below Q1. And Q3 is the 75\textsuperscript{th} percentile, because 75\% of the data falls below Q3.

• We typically don’t define a 100\textsuperscript{th} percentile, because there is no value for which 100\% of the data falls below it. After all, a number cannot be bigger than itself.

(This is why, on standardized tests, the 99\textsuperscript{th} percentile is the highest score possible)
Percentiles

Visually, the 10\textsuperscript{th} percentile is the location on the horizontal line for which the area to its left equals 10%.
Reverse standard normal example:
Determine a $z$ value from a percentile

• What height separates the bottom 10% from the top 90% of U.S. adult male heights?

• i.e., What is the $10^{th}$ percentile of U.S. adult male heights?

• To find this, we need to find an “$a$” such that

\[ P(z < a) = 0.10 \]
Reverse standard normal example

• Note that this is just the reverse of what we did before. Instead of finding the probability associated with a z value, we are finding the z value associated with a probability.

• So, the solution to this problem will be like what we have already done, but “backwards”

• Your calculator has a function, also under DIST, called “invNorm”. This function takes an area as input and gives you the corresponding percentile as a standard normal z-score.
Reverse standard normal example

- We want the 10\textsuperscript{th} percentile for height, so enter: \texttt{invNorm(0.1)}

- This is the z-score that separates the bottom 10\% of the standard normal distribution from the top 90\%. Visually:
• Now we need to convert this z-score into an x value (in this case, x is height).

• Recall the formula for finding a z-score from an X value:

\[ z = \frac{x - \mu}{\sigma} \]

• Doing a little bit of algebra to solve for x gives us:

\[ x = \mu + z\sigma \]

• This is the formula to use if you have a z-score and you want to turn it back into the “original”, non-standardized units.
• So, plug in $z = -1.28$, $\mu = 70$, and $\sigma = 3$ to this “reverse” $z$ formula:

$$x = \mu + z\sigma$$

• This is the 10$^{th}$ percentile of U.S. adult male heights. In other words, it is the height for which 10% of men are shorter and 90% are taller.
Reverse standard normal example 2

• Finally, let’s find the height that separates the top 5% of U.S. adult men from the rest.

• Our invNorm function only gives us percentiles for which the area will be to the LEFT of the computed z-value. So we must first take the complement of the area we are interested in. Visually:
Reverse standard normal example 2

• So, to find the height that separates the top 5% of men from the rest, we need to find the 95\textsuperscript{th} percentile.

• Enter \texttt{invNorm(0.95)} into your calculator. Then use this z-score to compute the 95\textsuperscript{th} percentile for U.S. adult male height:
Conclusion

• Much of what we do in statistics involves calculating probabilities.

• Probability distributions tell us the values a variable can take on, and how likely it is that the variable will take on any given value or range of values.

• The normal distribution is the most commonly used probability distribution. If we know a variable is normally distributed, we can use the known properties of the normal distribution to calculate the probability this variable will take on certain values.

• Probability distributions will be used extensively when we study inferential statistics.