Continuous Distributions

Tom Bruning

2020-08-29

Introduction

This class is a review of continuous distributions. The normal distribution is the backbone of the material in this course, and without a solid understanding of normal distributions this material will be a real slough.

- A continuous random variable is a variable that can assume any value on a continuum (can assume an uncountable number of values)
- These can potentially take on any value depending only on the ability to precisely and accurately measure

Normal Distribution

The normal distribution is characterized by it being:

- Bell Shaped
- Symmetrical
- Mean, Median and Mode are Equal
  - The Location is determined by the mean, μ
  - The Spread is determined by the standard deviation, σ
  - The random variable has an infinite theoretical range: $-\infty \to +\infty$

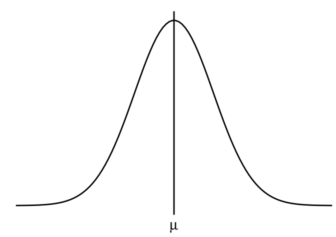

Figure 1: Generic Normal Distribution

Varying the parameters μ and σ

In Figure 2, we obtain different normal distributions

- type 1 and type 2 have the same means and different standard deviations and
- type 2 and type 3 have different means and different standard deviations

Changing μ shifts the distribution left or right.
Changing σ increases or decreases the spread.

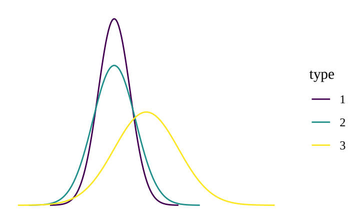

Figure 2: Comparing means and standard deviations

The Standardized Normal

Any normal distribution (with any mean and standard deviation combination) can be transformed into the standardized normal distribution (Z)

To compute normal probabilities we need to transform X units into Z units. The standardized normal distribution (Z) always has a mean of 0 and a standard deviation of 1.

$$Z = \frac{X - \mu}{\sigma} \tag{1}$$

To translate from X to the standardized normal (the "Z" distribution) we subtract the mean of X and dividing by its standard deviation. The Z distribution always has mean = 0 and standard deviation = 1.

For example: If X is distributed normally with mean of \$100 and standard deviation of \$50, the Z value for X = \$200 is:

$$Z = \frac{200 - 100}{50} = 2 \tag{2}$$

This says that X = \$200 is two standard deviations (2 increments of \$50 units) above the mean of \$100.

Empirical Rule

The emperical rule states:

- $\mu\pm$ 1σ covers about 68.26% of X's
- $\mu\pm$ 2σ covers about 95.44% of X's
- $\mu\pm$ 3σ covers about 99,73% of X's

Evaluating Normality

- Not all continuous distributions are normal
- It is important to evaluate how well the data set is approximated by a normal distribution.
- Normally distributed data should approximate the theoretical normal distribution:
- The normal distribution is bell shaped (symmetrical) where
  - the mean is equal to the median.
  - The empirical rule applies to the normal distribution.
  - The interquartile range of a normal distribution is 1.33 standard deviations.

Comparing data characteristics to theoretical properties

Construct charts or graphs

- For small- or moderate-sized data sets, construct a stem-and-leaf display or a boxplot to check for symmetry

- For large data sets, does the histogram or polygon appear bell-shaped?

  Compute descriptive summary measures
  * Do the mean, median and mode have similar values?
  * Is the interquartile range approximately 1.33 ?
  * Is the range approximately 6σ?
  Comparing data characteristics to theoretical properties

- Observe the distribution of the data set
  - Do approximately 2/3 of the observations lie within mean ± 1 standard deviation?
  - Do approximately 80% of the observations lie within mean ± 1.28 standard deviations?
  - Do approximately 95% of the observations lie within mean ± 2 standard deviations?
- Evaluate normal probability plot
  - Is the normal probability plot approximately linear (i.e. a straight line) with positive slope?

Bayes Theorem

Why This Is Important?
Suppose:

- You believe something is true, (~100% sure)
- And then some event happens which might make you think the original statement isn't true,
- How does this subsequent event affect your original belief?
  Now you might be less than 100% sure.

Bayes Theorem quantifies this type problem using conditional probabilities.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{3}$$

where A and B are events and P(B) $\neq$ 0.

- P(A) and P(B) are the probabilities of observing A and B without regard to each other.
- P(A | B), a conditional probability, is the probability of observing event A given that B is true.
- P(B | A) is the probability of observing event B given that A is true.

- P(A) is the prior probability (before Event B occurs), also called the baseline
- P(A|B) is the posterior probability (after Event B occurs), and can also the used as a new prior probability as subsequent Event Bs occur

Bayes Theorem Example

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

- Event A = Patient has the disease
- Event B = Test for the disease
- There is a disease that affects 0.1% of the population: P(A) = .001
- There is a test that
  - correctly identifies 100% of infected people as infected, and = P(B|A) called the *sensitivity* or true positive rate
  - correctly identifies 95% of uninfected people as uninfected = P(not B | not A) $P(\neg B|\neg A)$ called the *specificity* or true negative rate

Question: What is the probability that someone whom the test identifies as having the disease actually has the disease?