Probability, Gaussian Distribution and Z-scores

Evan Misshula

2017-02-16

Outline

Learning Objectives

Objectives for session 3

- Understand inferential statistics
- Explain the difference between parameter and a statistic
- Explain the difference between a population and a sample
- Probability review
- Introduction to the normal curve

Introduction

Inferential analysis

- Examine characteristics of a sample and make inferences about the population

Introduction

Inferential analysis

- Examine characteristics of a sample and make inferences about the population
- Collecting data on the entire population is too costly, too time consuming, and impossible

Inferential process

Our process is to make our decision with probability of error

- about whether sample characteristic is different from population characteristic

Definitions

Definition (Inferential statistics definitions)

- Population: entire group of study
- Sample: subset of population drawn to allow statistical analysis
- Parameter: a characteristic of the population
- Statistic: a characteristic of the sample
  - Used to make inferences to the population

Population vs. Sample

Population vs. Sample

- A population is the set of ALL the individuals of interest in a particular study

Population vs. Sample

- A population is the set of ALL the individuals of interest in a particular study
- Populations can vary in size from extremely large to very small
  - It depends on how the researcher defines the population in the study

Population examples

- Extremely large: All citizens in the United States
- Very small: All professors teaching statistics in JJ this year
- A population need NOT consist of people or individuals

Population examples

- Extremely large: All citizens in the United States
- Very small: All professors teaching statistics in JJ this year
- A population need NOT consist of people or individuals

- It can be anything you want to study a population of cats, dogs, rats, organizations (e.g., jails)

The sample

- Although we are always concerned with the entire population or start with a general question about the population, it is close to impossible to study ALL individuals in the population
- Typically, we select a small or manageable group of individuals from the targeted population for investigation
- This is what we called sampling or a sample
- A sample is a set of individuals selected from a population, usually intended to represent the population in a study

The sample

- Although we are always concerned with the entire population or start with a general question about the population, it is close to impossible to study ALL individuals in the population
- Typically, we select a small or manageable group of individuals from the targeted population for investigation
- This is what we called sampling or a sample
- A sample is a set of individuals selected from a population, usually intended to represent the population in a study

- Samples are expected to be "representative" of the populations
- A sample is a portion or a subset of the population

Probability

Can estimate probability of:

- Event occurring

Probability

Can estimate probability of:

- Event occurring
- Sample statistic matching population parameter

Probability

Can estimate probability of:

- Event occurring
- Sample statistic matching population parameter
- Number of times an event can occur divided by number of times any event can occur

Probability

Can estimate probability of:

- Event occurring
- Sample statistic matching population parameter
- Number of times an event can occur divided by number of times any event can occur
  - Important for inferential analysis because it represents probability of making a wrong decision about null hypothesis

Probability

Can estimate probability of:

- Event occurring
- Sample statistic matching population parameter
- Number of times an event can occur divided by number of times any event can occur
  - Important for inferential analysis because it represents probability of making a wrong decision about null hypothesis
- Probability
- Range: "0" to "1"
  - “0” = impossible
  - “1” = certain

Probability Laws

Theorem (Laws of probability (independent events))

- *Probability of one event OR another event is the sum of their probabilities*

Probability Laws

Theorem (Laws of probability (independent events))

- *Probability of one event OR another event is the sum of their probabilities*
- *Probability of one event AND another event is the product of their probabilities*

Gaussian Distribution

Gaussian also called Normal Distribution

- The only distribution completely described by two non-degenerate parameters, the mean and standard deviation
- Bell Shaped ("Bell Curve")
- Mean = Median = Mode
- Unimodal
- exponentially declining tails (big deviations are increasingly unlikely)

Gaussian distribution

Gaussian distribution probabilities

- 68% is within one standard deviation
- 95% are within two standard deviations
- 99% are within three standard deviations

Gaussian distribution

Gaussian distribution probabilities

- 68% is within one standard deviation
- 95% are within two standard deviations
- 99% are within three standard deviations
  - This only holds for Gaussian Distributions

Working with Gaussians

It is useful to translate

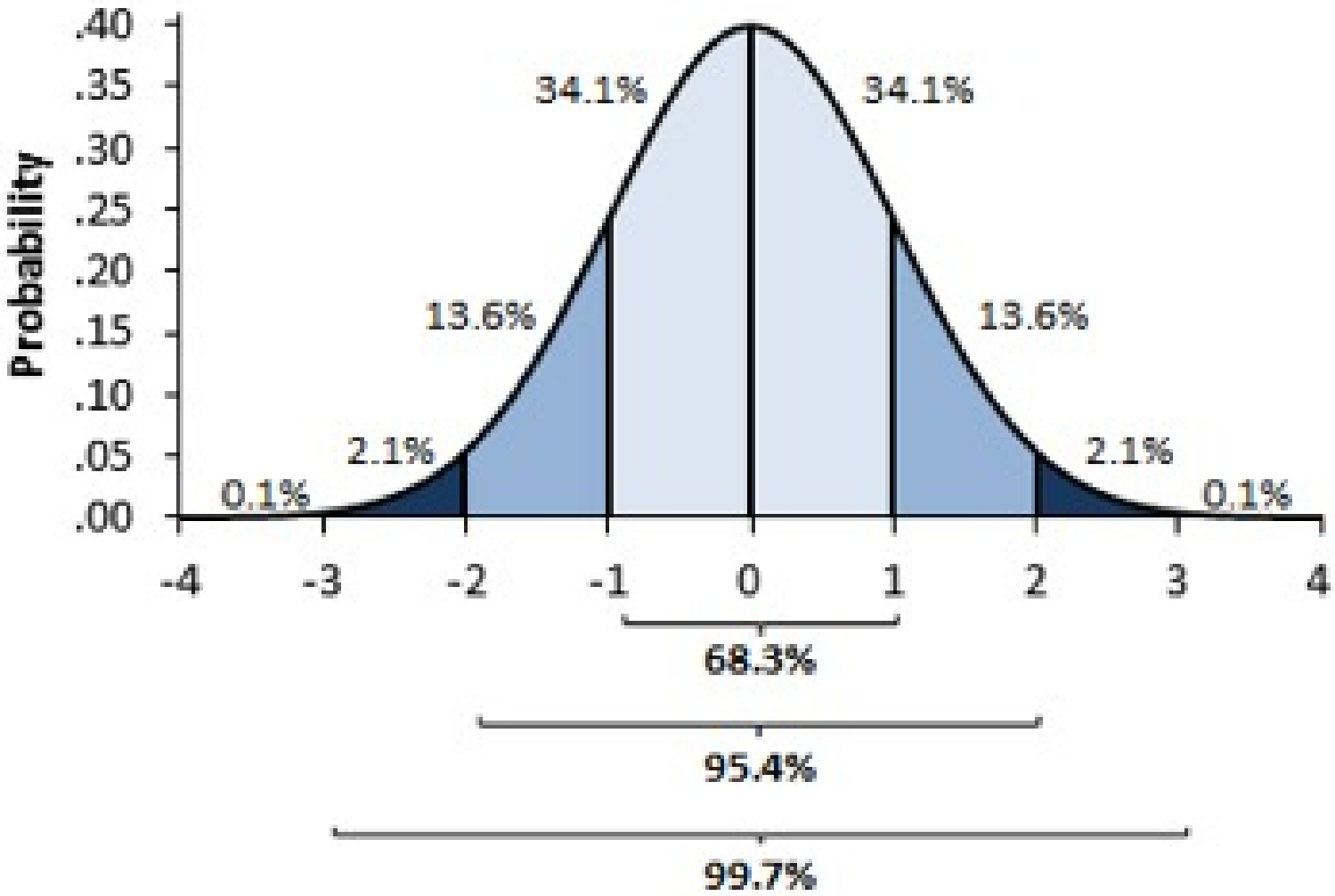
- all normal distributions of raw scores into a standard normal distribution Specifically, the Z distribution.
- The Z distribution standardizes the mean with a value of 0 and plots standard deviations in units of 1
- You can do this with a variable that is approximately normal

Why we care?

The standard normal distribution

- curve is extremely important in inferential statistics in the sense that it serves as a probability model for answering a lot of probability questions so that appropriate generalizations about populations can be made.
- In descriptive statistics we use it to give individuals a measure of position compared to others.
- It is because the Z distribution can be described or defined by the proportions of the area contained in each section of the distribution, NOT JUST 1,2, 3 & -1, -2, -3 std. deviations.
- The total area taken up by 1, 2,3 and 4 std. deviations from the mean. Total area is equal to 100%. 100% of the distribution falls within a Z score of +4 through -4.

Normal Curve



Z score example

Z score problems

- The mean commute time is: 34 minutes
- The standard deviation is: 7 minutes
  1. If an individual has a commute time that is two standard deviations above the mean, what is his commute time?
  2. How many standard deviations away from the mean is someone with a commute time of 27 minutes?
  3. What percent of the class has a commute time above that 27 minutes?

Two SD above the mean

What is his commute time?

```
commuteMean <- 34
commuteSD <- 7
TwoSDaboveMean <- commuteMean+2*commuteSD
TwoSDaboveMean

[1] 48
```

How many standard deviations away from the mean is someone with

- a commute time of 27 minutes?

```
Zscore27 <- (27-commuteMean)/commuteSD
Zscore27
```

```
[1] -1
```

What percent of the class has a commute time above that 27

- minutes?

```
pnorm(Zscore27,mean=0,sd=1,lower.tail = F)
Zscore27

[1] 0.8413447
[1] -1
```

Z score formula

Definition (The formula for a Z-score is)

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

Z score uses

1. Comparing scores from distributions with different means and standard deviations. (Such as scores on two different tests)
2. Finding the probability that a score will be located in a particular area. (Such as between 1.5 and 2.0 standard deviations)
3. Finding the probability that a score will be above a particular point. (Such as above .50 standard deviations)
4. Finding the probability that a score will be below a particular point. (Such as below 1.79 standard deviations)

Z scores and Probability

The area under the normal curve represents the measure of probability.

Z scores and Probability

The area under the normal curve represents the measure of probability.

- 90% of the area = a 90% (or .9000) probability.

Z scores and Probability

The area under the normal curve represents the measure of probability.

- 90% of the area = a 90% (or .9000) probability.
- The total area under the curve = 1.00 or 100%.

Z scores and Probability

The area under the normal curve represents the measure of probability.

- 90% of the area = a 90% (or .9000) probability.
- The total area under the curve = 1.00 or 100%.
- The probability of being somewhere under the curve is 1.

Z scores and Probability

The area under the normal curve represents the measure of probability.

- 90% of the area = a 90% (or .9000) probability.
- The total area under the curve = 1.00 or 100%.
- The probability of being somewhere under the curve is 1.
- The probability of not being somewhere under the curve is 0.

Z scores and Probability

The area under the normal curve represents the measure of probability.

- 90% of the area = a 90% (or .9000) probability.
- The total area under the curve = 1.00 or 100%.
- The probability of being somewhere under the curve is 1.
- The probability of not being somewhere under the curve is 0.
- Z scores can be negative, but probabilities cannot.

Z score problems

A population of adult heights is a

- Gaussian variable with a $\mu = 68$ inches
- and the population std dev $\sigma = 6$ inches

Z score problems

A population of adult heights is a

- Gaussian variable with a $\mu = 68$ inches
- and the population std dev $\sigma = 6$ inches

1. What is the probability of randomly selecting an individual from this population who is taller than 80 inches or p (X > 80)?

Z score method

When asked to find probabilities using raw scores

- that are normally distributed:

Z score method

When asked to find probabilities using raw scores

- that are normally distributed:
- The raw score must be turned into a z score.

The steps

Using the following procedure to find out the answer

1. Identify the exact position of $X = 80$ by computing a Z-score

```
Xbar<-68
sigma<-6
X<-80
Z<- (X-Xbar)/sigma
pnorm(Z, mean=0, sd=1, lower.tail=F)

[1] 0.02275013
```

Probabilities of Z scores

We can use the the R function, 'pnorm' to find the probability

- of getting scores between/beyond & below 1, 2 and 3 std. deviations or more accurately states between/beyond & below Z scores of positive & negative 1,2 & 3.

Z score exercise

What if I want to know: p (Z>1.5) p (Z<1.5) p (Z <.2) P (Z > .34)

```
pnorm(1.5, mean=0, sd=1, lower.tail=F) # p(Z>1.5)
pnorm(1.5, mean=0, sd=1, lower.tail=T) # p(Z<1.5)
pnorm(0.2, mean=0, sd=1, lower.tail=F) # p(Z>.2)
pnorm(0.34, mean=0, sd=1, lower.tail=F) # p(Z>.34)
```

```
[1] 0.0668072
[1] 0.9331928
[1] 0.4207403
[1] 0.3669283
```

Using the 'pnorm' function

Using the 'pnorm' function

1. Sketch the normal curve. Plot the mean and the Z score you are being asked about.

Using the 'pnorm' function

1. Sketch the normal curve. Plot the mean and the Z score you are being asked about.

1. Shade the area you need to find. If it is to 'the left' then lower tail is true (T), otherwise false (F).

In class exercise

Given =100 and =15, what is p (X<130)?

In class exercise

Given =100 and =15, what is p (X<130)?

```
myX <- 130
Xbar <- 100
mySigma <- 15
myZ <- (myX-Xbar)/mySigma
pnorm(myZ, mean=0, sd=1, lower.tail=T) # p(X<130)

[1] 0.9772499
```

Another in class exercise

Officer arrests are normally distributed with a μ= 73 and a

- standard deviation of 1.2. What is the probability of an officer having fewer than 71 arrests?

Another in class exercise

Officer arrests are normally distributed with a μ= 73 and a

- standard deviation of 1.2. What is the probability of an officer having fewer than 71 arrests?

```
myX <- 71
Xbar <- 73
mySigma <- 1.2
myZ <- (myX-Xbar)/mySigma
pnorm(myZ, mean=0, sd=1, lower.tail=T) # p(X<71)

[1] 0.04779035
```