Inferential Statistics

Week 08 Day 02

DS 3000 – Foundations of Data Science

1

Reminders

FP2: Datasets

Graded

HW 5

Motion to Dismiss

2

Outline

More on Normal Distribution

z-Scores

Probability Distributions

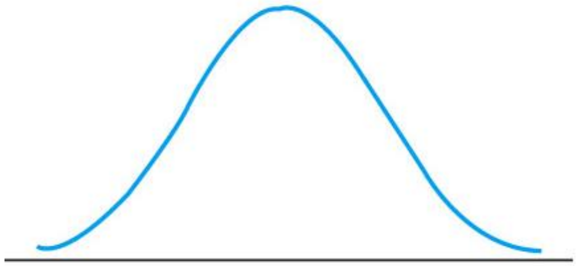
Standard Error

3

The Normal Distribution

In an ideal world our data would be distributed symmetrically around the center of all scores

This is known as a normal distribution and is characterized by the bell-shaped curve



4

Characteristics of the Normal Distribution

Symmetrical distribution

Each side is a mirror image of the other

The mean, median, and mode are all in the same place, in the center of the distribution

Theoretical distribution of the scores in the population

One rarely, if ever, gets a distribution of scores from a sample that forms an exact, normal distribution

Rather we hope to approach a normal, bell-shaped curve

Crucial for inferential statistics

5

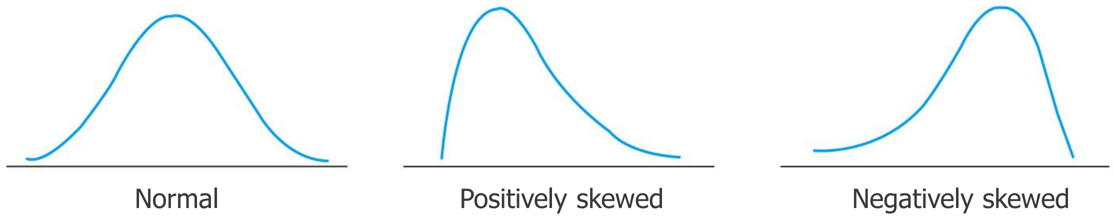
Skewed Distributions

Skewness denotes lack of symmetry

Scores pile up on one side and taper off in a **tail** on the other

Tail on the right (high scores) = positive skew

Tail on the left (low scores) = negative skew



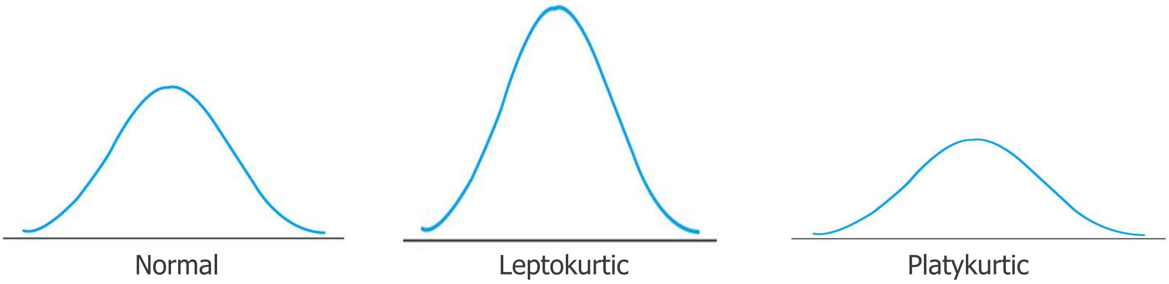
6

Kurtosis

Refers to the shape of the distribution in terms of height, or flatness

Leptokurtic distributions = more pointy (higher peak)

Platykurtic distributions = flatter (lower peak)



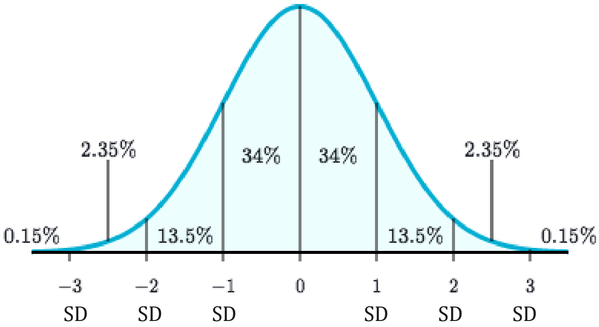
7

The Empirical Rule

The 68-95-99.7 Rule

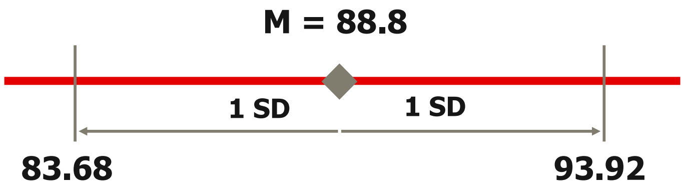
For a normal distribution, nearly all of the data will fall within three standard deviations of the mean.

- 68% within 1 SD of the mean
- 95% within 2 SD of the mean
- 99.7% within 3 SD of the mean



8

The Mean and Standard Deviation



M = 88.8

1 SD

1 SD

83.68

93.92

The distribution of DADA scores:

80, 82, **86**, **87**, **89**, **90**, **92**, **92**, 94, 96

z-Scores

The Tale of Two Scores in a Distribution

How can you interpret the performance of each student relative to the mean without knowing the mean?

Neville = 90

Ron = 94

11

z-Scores

Identify and describe location of every score in the distribution

Exact location is described by *z*-score

$$z = \frac{x - \bar{x}}{SD}$$

Sign tells whether score is located above or below the mean

Number tells distance between score and mean in standard deviation units

12

z-Scores

Standardize an entire distribution

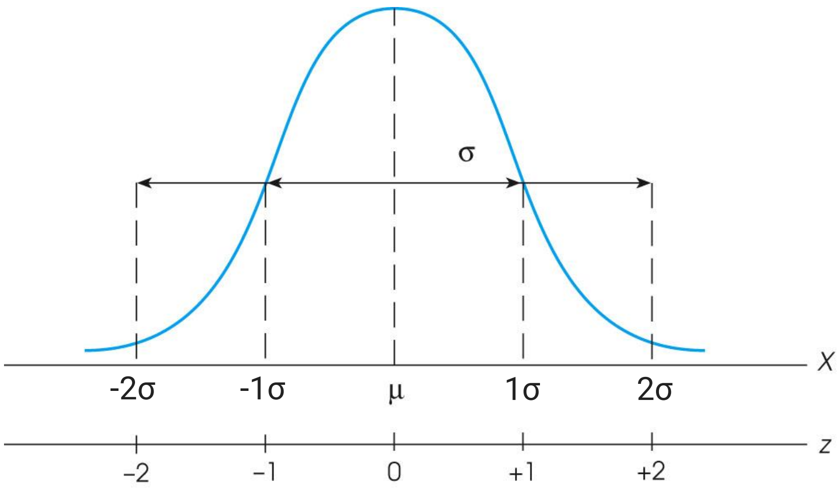
The mean of the distribution becomes zero

The standard deviation becomes 1

Take different distributions and make them equivalent and comparable

Can compare two scores from a distribution in standard deviation units

z-Scores and Location in a Distribution



Computing z-Scores

The mean DADA scores was 88.8 (SD = 5.12)

For Ron (x = 94),

$$z = \frac{94 - 88.8}{5.12}$$

$$z = 1.02$$

For Neville (x = 90),

$$z = \frac{90 - 88.8}{5.12}$$

$$z = 0.23$$

A z-score of 1.02 indicates that Ron's score is 1.02 SDs above the mean

A z-score of 0.23 indicates that Neville's score is 0.23 SDs above the mean

15

Comparing Two Distributions

Suppose Neville took two exams, Charms and Transfiguration, and got 76 on both.

Given the following info about the exams, on which test did he do better?

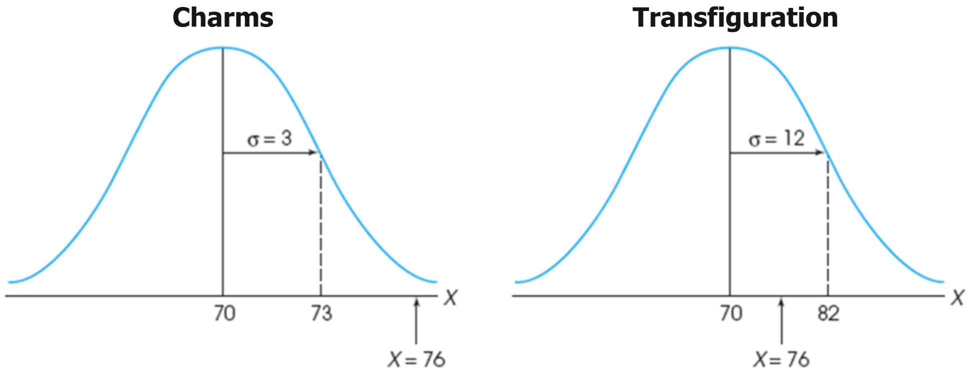
Charms	**Transfiguration**
Average = 70	Average = 70
SD = 3	SD = 12

16

Comparing Two Distributions

Suppose Neville took two exams, Charms and Transfiguration, and got 76 on both.

Given the following info about the exams, on which test did he do better?



Charms

Transfiguration

Comparing Two Distributions

Simply compute the z-score for each test and compare the z-scores

Neville’s **Charms** score (x = 76):

$z = (76-70)/3 = 2$

Neville’s **Transfiguration** score (x = 76):

$z = (76-70)/12 = .5$

Neville’s performance on the exams:

z = 2 SDs above the mean on Charms

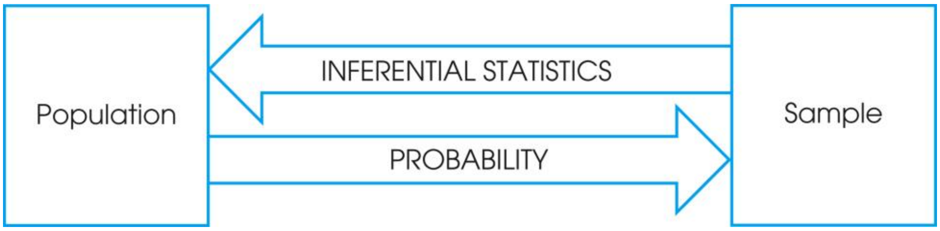
z = .5 SDs above the mean on Transfiguration

Probability

Role of Probability in Inferential Statistics

Inferential statistics use sample data to answer questions about the population

Relationships between samples and populations are defined in terms of probability



Defining Probability

Probability is the likelihood or chance that some random event will occur.

It is described by

$$\boxed{0 < P(E) < 1}$$

0 means that there is no chance that the event happens

1 means that the event happens with certainty.

Defining Probability

If an event has a chance P of occurring then there’s a chance of 1-P that it will not occur.

$$\boxed{P(E) = 1 - P(E')}$$

Defining Probability

Laplace (1774)

The ratio of the number of desirable outcomes to the total number of possible outcomes, provided all outcomes are equally likely

$$\textit{probability of } A = \frac{\textit{number of outcomes classified as } A}{\textit{total number of possible outcomes}}$$

Probability Notation

p is the symbol for “probability”

Probability of some specific outcome is specified by *p*(event)

So the probability of rolling a 5 on a die could be symbolized as ***P*****(one)**

P(one) = 1/6 ≈ 0.1667 (proportion is 1 face out of 6 faces)

16.67% chance of rolling a 5 on a die (or any other face)

Probabilities are always proportions

Random Sampling

A process or procedure used to draw samples

Required for our definition of probability to be accurate

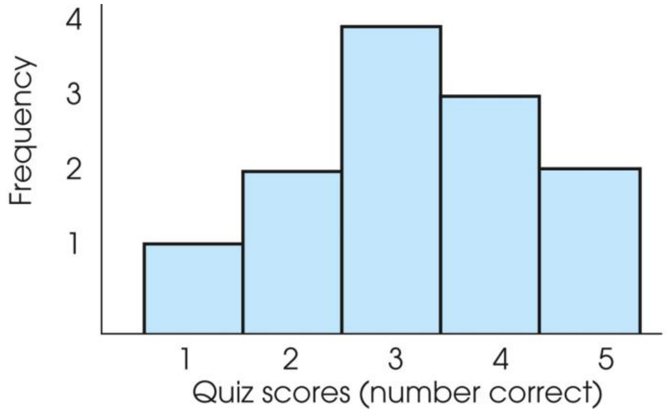
A random sample is produced by a process that assures:

- Each data point in the population has an equal chance of being selected
- Probability of being selected stays constant from one selection to the next when more than one data point is selected

Requires sampling with replacement

25

Probability and Frequency Distributions



X	f
5	2
4	3
3	4
2	2
1	1

How likely is it that a student scored 5 on the quiz?
What is the probability of a student scoring 5 on the quiz?

26

Question

Professor Lupin's class has 48 students with an equal number of students from each of the four houses of Hogwarts (Gryffindor, Hufflepuff, Ravenclaw, and Slytherin).

If Professor Lupin randomly selects a student to perform the Riddiculus Spell, what is the probability of selecting a student from Gryffindor?

This is a fair selection. No cheating or anything like that.

27

Probability and Frequency Distributions

Probability usually involves population of scores that can be displayed in a frequency distribution graph, or histogram

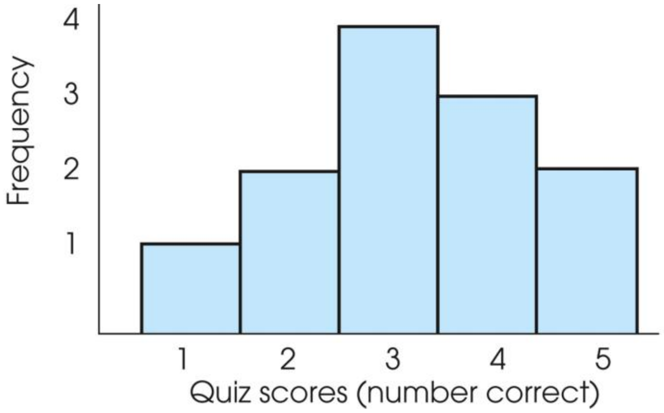
A particular portion of the graph corresponds to a particular probability in the population

The size of the bars relate directly to the probability of an event occurring

Proportions and probabilities are equivalent

28

Probability and Frequency Distributions



X	f
5	2
4	3
3	4
2	2
1	1

How likely is it that a student scored 5 on the quiz?

What is the probability of a student scoring 5 on the quiz?

29

Probability and Frequency Distributions

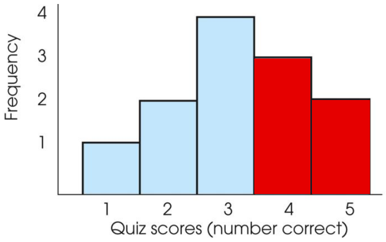
How likely is it that a student scored higher than 3 on the quiz?

How big is the red area of the histogram compared to the total size of all bars?

To find the size of the red region, add the values of the bars: 3 + 2 = 5

Total size of all bars: 12

P(score>3) = 5 / 12 = 0.4167



X	f
5	2
4	3
3	4
2	2
1	1

30

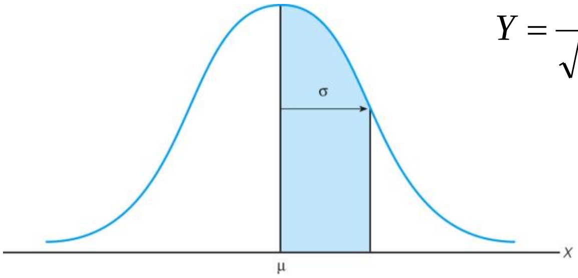
Probability and the Normal Distribution

Normal distribution is a common shape

Defined by an equation

Can be described by the proportions of area contained in each section

z-scores are used to identify sections



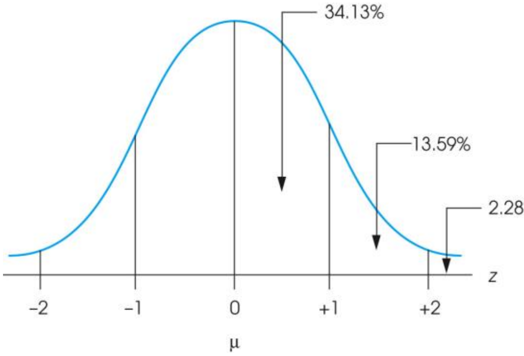
$$Y = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(X-\mu)^2/2\sigma^2}$$

Probability Distributions

Because z-scores define the sections, the proportions of area apply to any normal distribution

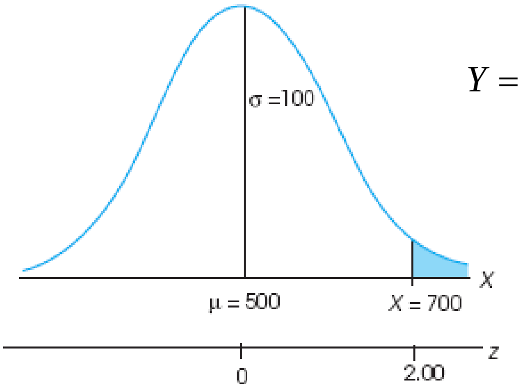
Regardless of the mean

Regardless of the standard deviation



Probability Distributions

The distribution of SAT scores



$$Y = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(X-\mu)^2/2\sigma^2}$$

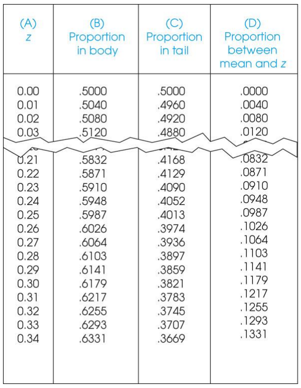
The blue shaded region corresponds to the probability of an SAT score being 700 or greater

33

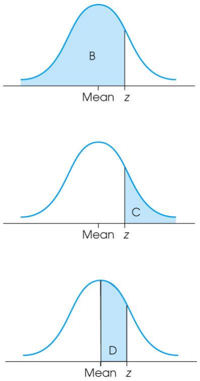
Probability Distributions

Probability values, area under the curve, for all possible z-score values have been calculated by statisticians

Available as z Score Table, Standard Normal Table, or Unit Normal Table



(A) z	(B) Proportion in body	(C) Proportion in tail	(D) Proportion between mean and z
0.00	.5000	.5000	.0000
0.01	.5040	.4960	.0040
0.02	.5080	.4920	.0080
0.03	.5120	.4880	.0120
0.21	.5832	.4168	.0832
0.22	.5871	.4129	.0871
0.23	.5910	.4090	.0910
0.24	.5948	.4052	.0948
0.25	.5987	.4013	.0987
0.26	.6026	.3974	.1026
0.27	.6064	.3936	.1064
0.28	.6103	.3897	.1103
0.29	.6141	.3859	.1141
0.30	.6179	.3821	.1179
0.31	.6217	.3783	.1217
0.32	.6255	.3745	.1255
0.33	.6293	.3707	.1293
0.34	.6331	.3669	.1331



34

Probability Distributions

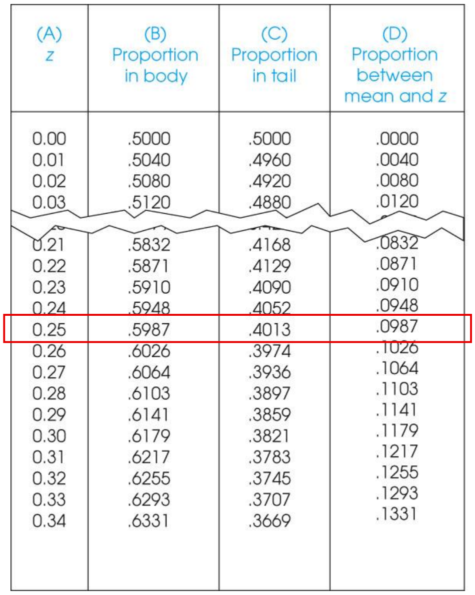
z = .25

P(z or greater) =

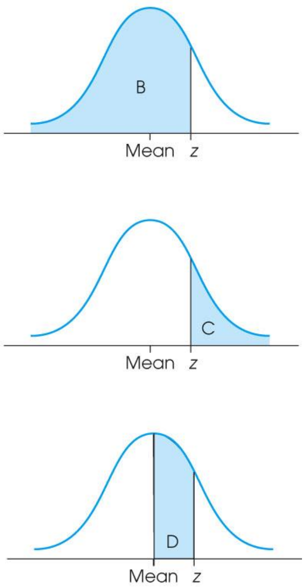
.4013

P(below z) =

.5987



(A) z	(B) Proportion in body	(C) Proportion in tail	(D) Proportion between mean and z
0.00	.5000	.5000	.0000
0.01	.5040	.4960	.0040
0.02	.5080	.4920	.0080
0.03	.5120	.4880	.0120
0.21	.5832	.4168	.0832
0.22	.5871	.4129	.0871
0.23	.5910	.4090	.0910
0.24	.5948	.4052	.0948
0.25	.5987	.4013	.0987
0.26	.6026	.3974	.1026
0.27	.6064	.3936	.1064
0.28	.6103	.3897	.1103
0.29	.6141	.3859	.1141
0.30	.6179	.3821	.1179
0.31	.6217	.3783	.1217
0.32	.6255	.3745	.1255
0.33	.6293	.3707	.1293
0.34	.6331	.3669	.1331

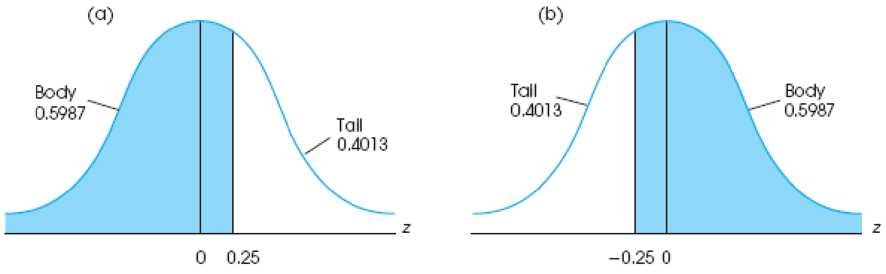


35

Probability Distributions

Proportions of a normal distribution corresponding to

z = +0.25 (a) and -0.25 (b).



36

Probability Distributions

Unit normal table lists relationships between *z*-score locations and proportions in a normal distribution

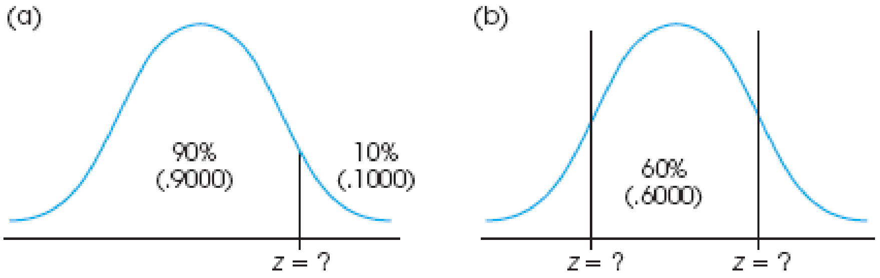
If you know the *z*-score, you can look up the corresponding proportion

If you know the proportion, you can use the table to find a specific *z*-score location

Probability is equivalent to proportion

37

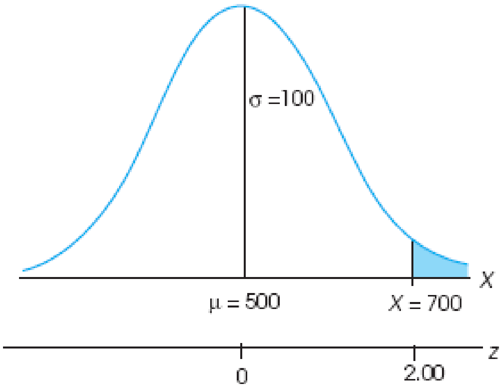
Probability Distributions



38

Probability Distributions

The distribution of SAT scores



What is the range of SAT scores between which the **middle** 95% of scores fall?

39

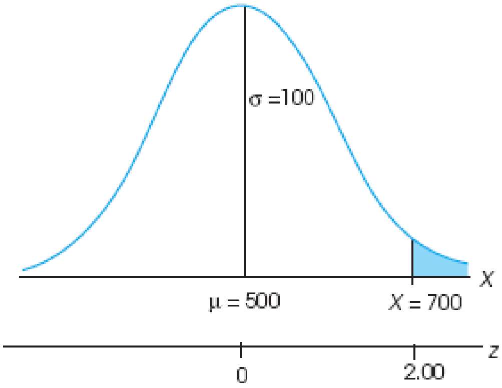
Probability Distributions

To get the middle 95%

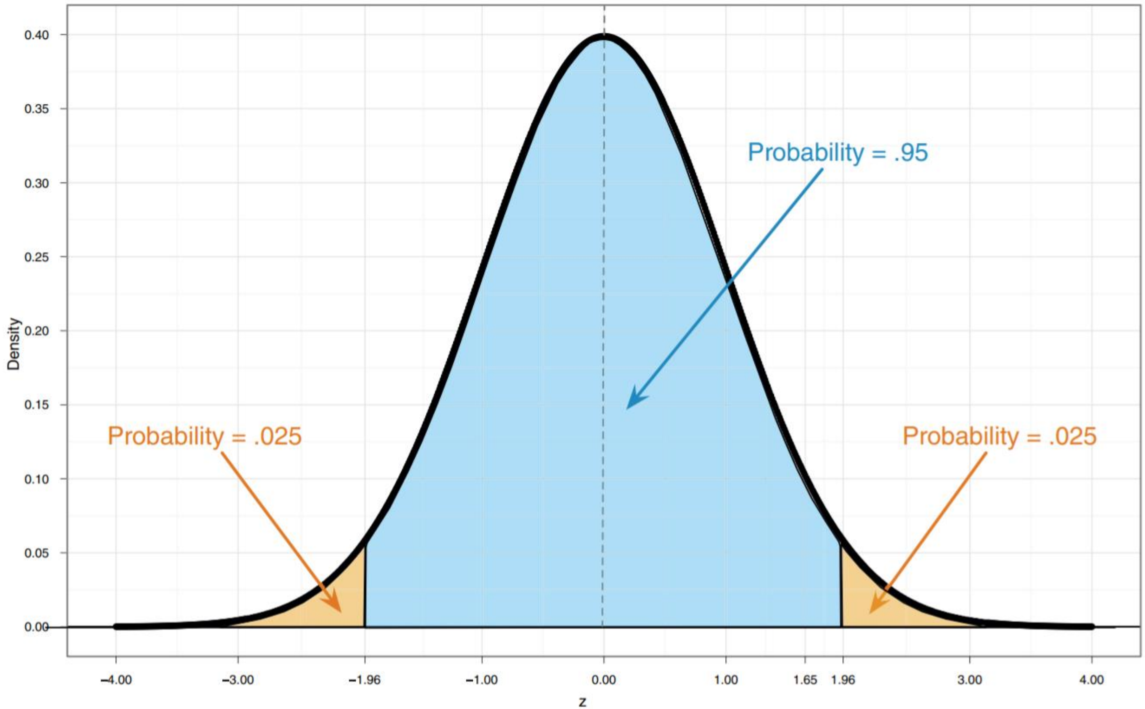
Need to cut off 2.5% of the scores on each end

Then find the z value that cuts off the top/bottom area of .025 in the table

The distribution of SAT scores



40



41

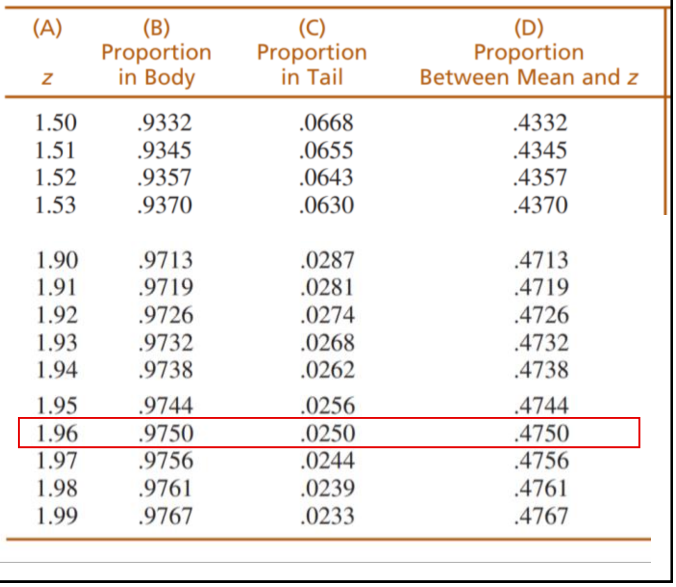
Probability Distributions

z = +/- 1.96 cuts off

top/bottom 2.5% of the scores

The middle 95% of z-scores fall between -1.96 and 1.96

Remember this!



(A) *z*	(B) Proportion in Body	(C) Proportion in Tail	(D) Proportion Between Mean and *z*
1.50	.9332	.0668	.4332
1.51	.9345	.0655	.4345
1.52	.9357	.0643	.4357
1.53	.9370	.0630	.4370
1.90	.9713	.0287	.4713
1.91	.9719	.0281	.4719
1.92	.9726	.0274	.4726
1.93	.9732	.0268	.4732
1.94	.9738	.0262	.4738
1.95	.9744	.0256	.4744
1.96	.9750	.0250	.4750
1.97	.9756	.0244	.4756
1.98	.9761	.0239	.4761
1.99	.9767	.0233	.4767

42

Probability Distributions

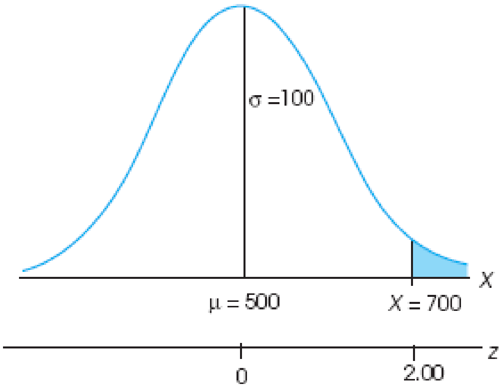
$$z = \frac{x - \bar{x}}{SD}$$

Upper = 1.96 * 100 + 500 = 696

Lower = -1.96 * 100 + 500 = 304

95% of the scores are between 304 and 696

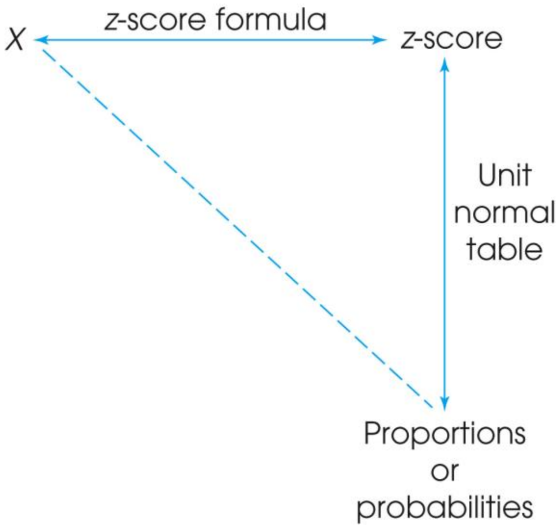
The distribution of SAT scores



How likely is it that a student will have an SAT score >= 696?

43

Determining Probabilities of Scores



X

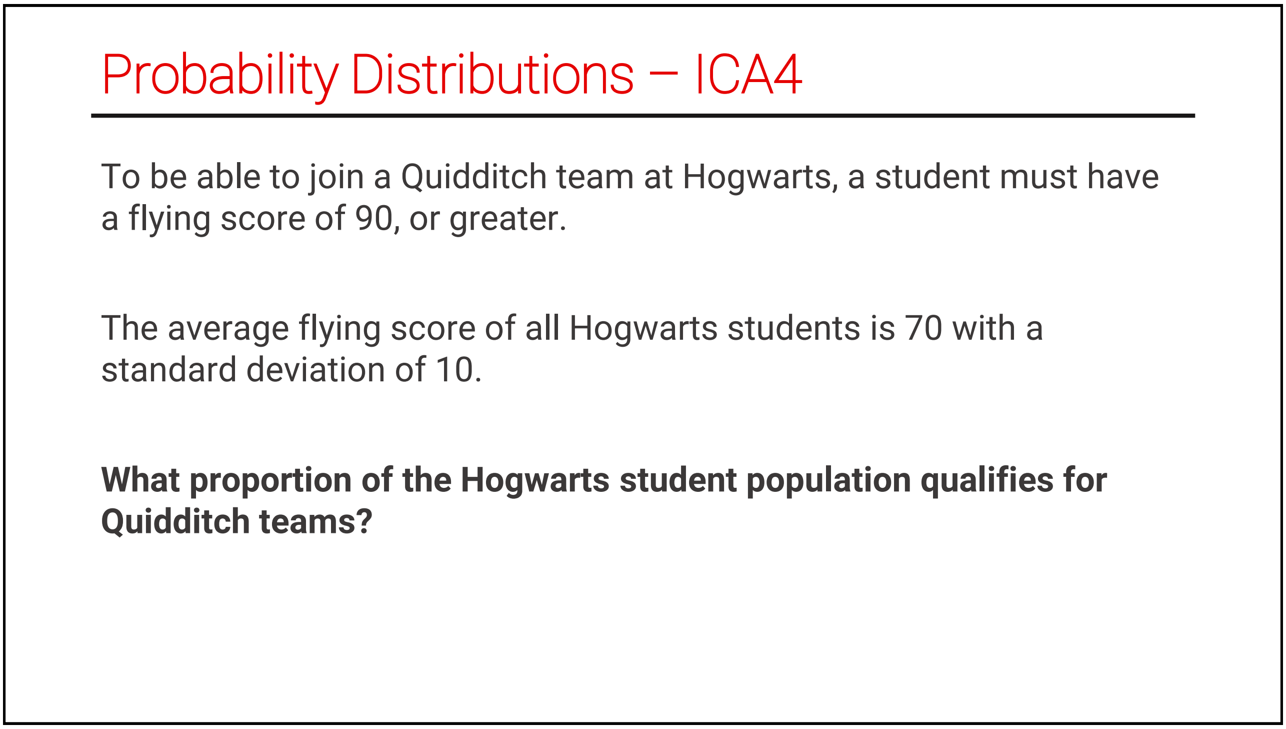
z-score formula

z-score

Unit normal table

Proportions or probabilities

44



Probability Distributions – ICA4

To be able to join a Quidditch team at Hogwarts, a student must have a flying score of 90, or greater.

The average flying score of all Hogwarts students is 70 with a standard deviation of 10.

What proportion of the Hogwarts student population qualifies for Quidditch teams?