Statistics for Computing

MA4413 Lecture 5A/5B

Kevin O'Brien

kevin.obrien@ul.ie

Dept. of Mathematics & Statistics,
University *of* Limerick

Autumn 2013

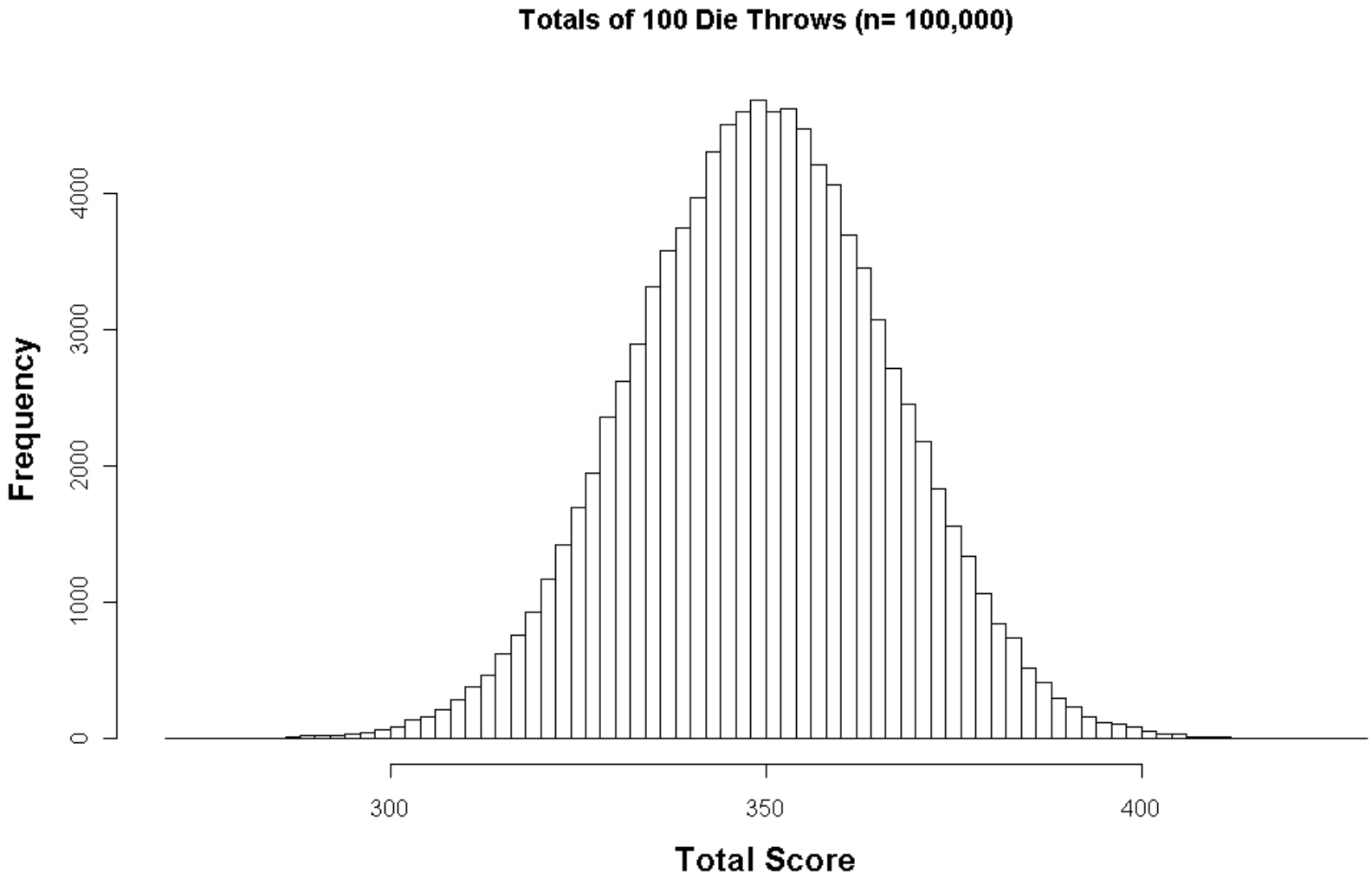
Lectures and Mid-Terms

- Lecture Slot 5A was used for the first Mid-term exam.
- The Lecture Series recommences in Lecture Slot 5B.
- Lecture Slot 4B covered the Exponential Distribution.
- The next mid-term is provisionally scheduled take place in Week 9. The normal distribution, hypothesis testing (including the t-test), confidence intervals, and inference procedures are the only topics going to be examined in this mid-term.
- Please be advised of sample papers for the second midterm in the SULIS directory.

Introduction to the Normal Distribution

- Recall the experiment whereby a die was rolled 100 times, and the sum of the 100 values was recorded.
- This experiment was repeated a very large number of times (e.g. 100,000 times) in a simulation study.
- A histogram was drawn to depict the distribution of outcomes of this experiment.
- Recall that we agreed that “bell-shaped” was a good description of the histogram.

Normal Distribution



Normal Distribution

- The normal distribution is perhaps the most widely used type of probability distribution for a random variable.
- Normal distributions have the same general shape: the bell curve.
- The distributions are **symmetric** with values concentrated more in the middle than in the tails.
- Important The height of a normal distribution can be defined mathematically in terms of two fundamental parameters: the normal mean (μ) and the normal standard deviation (σ).
- A normally distributed random variable X is denoted $X \sim \mathrm{N}(\mu, \sigma^2)$ (note that we use the variance term here).
- The mean (μ) and standard deviation (σ) are vital for calculating probabilities.

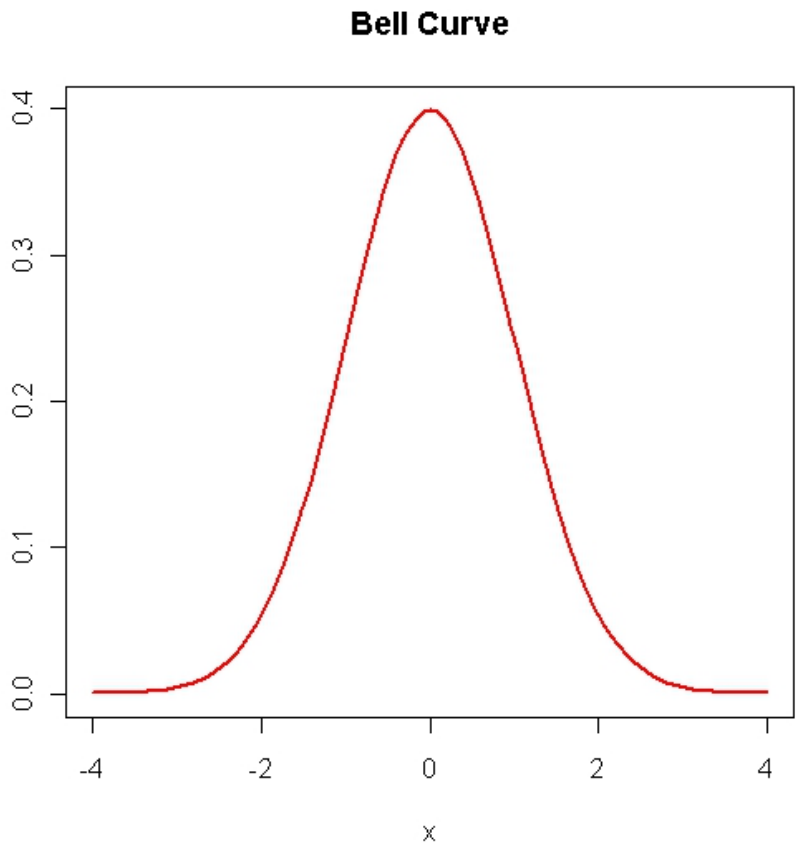
The Normal Distribution

The ***probability density function*** of the normal distribution is given as

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

Integrating this formula would allow us to compute probabilities.
However, it is not required to use this formula.

Normal Distribution



Characteristics of the Normal probability distribution

1. The highest point on the normal curve is at the mean, which is also the median of the distribution.
2. [VERY IMPORTANT] The normal probability curve is bell-shaped and symmetric, with the shape of the curve to the left of the mean a mirror image of the shape of the curve to the right of the mean. (This is the basis of an important rule, called the **Symmetry Rule**, that we shall meet later.)
3. The standard deviation determines the width of the curve. Larger values of the the standard deviation result in wider flatter curves, showing more dispersion in data.
4. As with all density curves, the total area under the curve for the normal probability distribution is 1.

Characteristics of the Normal probability distribution

Remark: It is useful to know the following statements as rules of thumb, but we will do all relevant calculations from first principles. However, in an exam situation, these rules of thumb may be invoked, and it is **NOT** required to show your workings.

- The interval defined by the mean $\pm 1\times$ standard deviation includes approximately 68% of the observations, leaving 16% (approx) in each tail.
- The interval defined by the mean $\pm 1.645\times$ standard deviation includes approximately 90% of the observations, leaving 5% (approx) in each tail.
- The interval defined by the mean $\pm 1.96\times$ standard deviation includes approximately 95% of the observations, leaving 2.5% (approx) in each tail.
- The interval defined by the mean $\pm 2.58\times$ standard deviation includes approximately 99% of the observations, leaving 0.5% (approx) in each tail.

The Standard Normal Distribution

- The standard normal distribution is a special case of the normal distribution with a mean $\mu = 0$ and a standard deviation $\sigma = 1$.
- We denote the standard normal random variable as Z rather than X.

$$Z \sim N(0, 1^2)$$

- The distribution is well described in statistical tables (i.e. Murdoch Barnes Table 3, aka MB3)
- Rather than computing probabilities from first principles, which is very difficult, probabilities from distributions other than the Z distribution (e.g. X $\sim$N($\mu = 100, \sigma = 15$)) can be computed using the Z distribution, a much easier approach. (We shall demonstrate how to do this shortly.)

Standardization formula

All normally distributed random variables have corresponding Z values, called ***Z-scores***.

For normally distributed random variables, the z-score can be found using the ***standardization formula***;

$$z_o = \frac{x_o - \mu}{\sigma}$$

where x_o is a score from the underlying normal ("X") distribution, μ is the mean of the original normal distribution, and σ is the standard deviation of original normal distribution.

Therefore z_o is the z-score that corresponds to x_o.

- Terms with subscripts mean particular values, and are not variable names.
- A computed Z-score is a normally distributed random variable only if the underlying distribution (X) is normally distributed. If the underlying distribution is not normal, then using Z-scores is not a valid approach.

The Standardized Value

- Suppose that mean $\mu = 80$ and that standard deviation $\sigma = 8$.
- What is the Z-score for $x_o = 100$?

$$z_{100} = \frac{x_0 - \mu}{\sigma} = \frac{100 - 80}{8} = \frac{20}{8} = 2.5$$

- Therefore the Z score is : $z_{100} = 2.5$

Z-scores

- A Z-score always reflects the number of standard deviations above or below the mean a particular score is.
- Suppose the scores of a test are normally distributed with a mean of 50 and a standard deviation of 9
- For instance, if a person scored a 68 on a test, then they scored 2 standard deviations above the mean.
- Converting the test scores to z scores, an X value of 68 would yield:

$$Z = \frac{68 - 50}{9} = 2$$

- So, a Z score of 2 means the original score was 2 standard deviations above the mean.

The Standard Normal (Z) Distribution Tables

- Importantly, probabilities relating to the z distribution are comprehensively tabulated in ***Murdoch Barnes Table 3***.
- This is available on sulis, in the "about this module" folder.
- Given a value of k (with k usually between 0 and 4), the probability of a standard normal "Z" random variable being greater than (or equal to) k $P(Z \geq k)$ is given in Murdoch Barnes table 3 .
- Other statistical tables can be used (e.g. the Dept. of Education Tables that many student would have used in school), but they may tabulate probabilities in a different way.

An Important Identity

If two values z_o and x_o are related in the following way, for some values μ and σ,

$$z_0 = \frac{x_0 - \mu}{\sigma}$$

Then we can can say

$$P(X \geq x_o) = P(Z \geq z_o)$$

or alternatively

$$P(X \leq x_o) = P(Z \leq z_o)$$

This is fundamental to solving problems involving normal distributions.

Using Murdoch Barnes Tables 3

- For some value z_o, between 0 and 4, the Murdoch Barnes tables set 3 tabulate $P(Z \geq z_o)$
- Ideally z_o would be specified to 2 decimal places. If it is not, round to the closest value.
- We call the third digit (i.e. the digit in the second decimal place) the "second precision".

Using Murdoch Barnes Tables 3

- To compute the relevant probability we express z_o as the sum of z_o without the second precision, and the second precision.(For example $1.28 = 1.2 + 0.08$.)
- Select the row that corresponds to z_o without the second precision (e.g. 1.2).
- Select the column that corresponds to the second precision(e.g. 0.08).
- The value that contained on the intersection is $P(Z \geq z_o)$

Find $P(Z \geq 1.28)$

	...	...	0.006	0.07	0.08	0.09
...	...	...	...	...	...	...
1.0	...	...	0.1446	0.1423	0.1401	0.1379
1.1	...	...	0.1230	0.1210	0.1190	0.1170
1.2	...	...	0.1038	0.1020	0.1003	0.0985
1.3	...	...	0.0869	0.0853	0.0838	0.0823
...	...	...	...	...	...	...

Using Murdoch Barnes tables 3

- Find $P(Z \geq 0.60)$
- Find $P(Z \geq 1.64)$
- Find $P(Z \geq 1.65)$
- Estimate $P(Z \geq 1.645)$

Find $P(Z \geq 0.60)$

	0.00	0.01	0.02	0.03	...	...
...	...	...	...	...	...	...
0.4	0.3446	0.3409	0.3372	0.3336	...	...
0.5	0.3085	0.3050	0.3015	0.2981	...	...
0.6	0.2743	0.2709	0.2676	0.2643	...	...
0.7	0.2420	0.2389	0.2358	0.2327	...	...
...	...	...	...	...	...	...

Find $P(Z \geq 1.64)$ and $P(Z \geq 1.65)$

	...	...	0.04	0.05	0.06	0.07
...	...	...	...	...	...	...
1.5	...	0.0630	0.0618	0.0606	0.0594	...
1.6	...	0.0516	0.0505	0.0495	0.0485	...
1.7	...	0.0418	0.0409	0.0401	0.0392	...
...	...	...	...	...	...	...

Using Murdoch Barnes tables 3

- $P(Z \geq 1.64) = 0.505$
- $P(Z \geq 1.65) = 0.495$

- $P(Z \geq 1.645)$ is approximately the average value of $P(Z \geq 1.64)$ and $P(Z \geq 1.65)$.
- $P(Z \geq 1.645) = (0.0495 + 0.0505)/2 = 0.0500$. (i.e. 5%)

Exact Probability

Remarks: This is for continuous distributions only.

- The probability that a continuous random variable will take an exact value is infinitely small. We will usually treat it as if it was zero.
- When we write probabilities for continuous random variables in mathematical notation, we often retain the equality component (i.e. the "...or equal to..").
  For example, we would write expressions $P(X \leq 2)$ or $P(X \geq 5)$.
- Because the probability of an exact value is almost zero, these two expression are equivalent to $P(X < 2)$ or $P(X > 5)$.
- The complement of $P(X \geq k)$ can be written as $P(X \leq k)$.

Complement and Symmetry Rules

Any normal distribution problem can be solved with some combination of the following rules.

- **Complement rule**
- Common to all continuous random variables

$$P(Z \geq k) = 1 - P(Z \leq k)$$

Similarly

$$P(X \geq k) = 1 - P(X \leq k)$$

$$P(Z \leq 1.28) = 1 - P(Z \geq 1.28) = 1 - 0.1003 = 0.8997$$

Complement and Symmetry Rules

- **Symmetry rule**
- This rule is based on the property of symmetry mentioned previously.
- Only the probabilities corresponding to values between 0 and 4 are tabulated in Murdoch Barnes.
- If we have a negative value of k, we can use the symmetry rule.

$$P(Z \leq -k) = P(Z \geq k)$$

by extension, we can say

$$P(Z \geq -k) = P(Z \leq k)$$

Z Scores: Example 1

Find $P(Z \geq -1.28)$

Solution

- Using the symmetry rule

$$P(Z \geq -1.28) = P(Z \leq 1.28)$$

- Using the complement rule

$$P(Z \geq -1.28) = 1 - P(Z \geq 1.28)$$

$$P(Z \geq -1.28) = 1 - 0.1003 = 0.8997$$

Z Scores: Example 2

Find the probability of a Z random variable being between -1.8 and 1.96? i.e. Compute $P(-1.8 \leq Z \leq 1.96)$
Solution

- Consider the complement event of being in this interval: a combination of being too low or too high.
- The probability of being too low for this interval is $P(Z \leq -1.80) = 0.0359$ (check)
- The probability of being too high for this interval is $P(Z \geq 1.96) = 0.0250$ (check)
- Therefore the probability of being **outside** the interval is 0.0359 + 0.0250 = 0.0609.
- Therefore the probability of being **inside** the interval is 1- 0.0609 = 0.9391

$$P(-1.8 \leq Z \leq 1.96) = 0.9391$$

Application : Example

The mean time spent waiting by customers before their queries are dealt with at an information centre is 10 minutes.
The waiting time is normally distributed with a standard deviation of 3 minutes.

i) What percentage of customers will be waiting longer than 15 minutes

ii) 90% of customers will be dealt with in at most 12 minutes. Is this statement true or false? Justify your answer.

iii) What percentage of customers will wait between 7 and 13 minutes before their query is dealt with?

Solutions

Let x be the normal random variable describing waiting times
$P(X \geq 15) = ?$

First , we find the z-value that corresponds to x = 15 (remember $\mu = 10$ and $\sigma = 3$)

$$z_o = \frac{x_o - \mu}{\sigma} = \frac{15 - 10}{3} = 1.666$$

- We will use $z_o = 1.67$
- Therefore we can say $P(X \geq 15) = P(Z \geq 1.67)$
- The Murdoch Barnes tables are tabulated to give $P(Z \geq z_o)$ for some value z_o .
- We can evaluate $P(Z \geq 1.67)$ as 0.0475.
- Necessarily $P(X \geq 15) = 0.0475$.

Solutions

- ”90% of customers will be dealt with in at most 12 minutes.”
- To answer this question, we need to know $P(X \leq 12)$
- First , we find the z-value that corresponds to x = 12 (remember $\mu = 10$ and $\sigma = 3$)

$$z_o = \frac{x_o - \mu}{\sigma} = \frac{12 - 10}{3} = 0.666$$

Solutions

- We will use $z_o = 0.67$ (although 0.66 would be fine too)
- Therefore we can say $P(X \geq 12) = P(Z \geq 0.67) = 0.2514$
- Necessarily $P(X \leq 12) = P(Z \leq 0.67) = 0.7486$
- 74.86% of customers will be dealt with in at most 12 minutes.
- The statement that 90% will be dealt with in at most 12 minutes is false.

Solutions

What percentage will wait between 7 and 13 minutes ?
$P(7 \leq X \leq 13) = ?$
Solution:
Compute the probability of being too low, and the probability of being too high for the interval.
The probability of being inside the interval is the complement of the combination of these events.

Solutions

Too high:

$P(X \geq 13) = ?$

$$z_o = \frac{13 - 10}{3} = 1$$

From tables, $P(Z \geq 1) = 0.1587$. Therefore $P(X \geq 13) = 0.1587$

Too low:

$P(X \leq 7) = ?$

$$z_o = \frac{7 - 10}{3} = -1$$

By symmetry, and using tables, $P(X \leq 7) = P(Z \leq -1) = 0.1587$

Solutions

$$P(7 \leq X \leq 13) = 1 - [P(X \leq 7) + P(X \geq 13)]$$

$$P(7 \leq X \leq 13) = 1 - [0.1587 + 0.1587] = 0.6826$$

Normal Distribution : Solving problems

Recap:

- We must know the normal mean μ and the normal standard deviation σ.
- The normal random variable is $X \sim \mathrm{N}(\mu, \sigma^2)$.
- (If we don't, we usually have to determine them, given the information in the question.)
- The standard normal random variable is $Z \sim \mathrm{N}(0, 1^2)$.
- The standard normal distribution is well described in Murdoch Barnes Table 3, which tabulates $P(Z \geq z_o)$ for a range of Z values.

Normal Distribution : Solving problems

- For the given value x_o from the variable X, we compute the corresponding z-score z_o.

$$z_o = \frac{x_o - \mu}{\sigma}$$

- When z_o corresponds to x_o, the following identity applies:

$$P(X \geq x_o) = P(Z \geq z_o)$$

- Alternatively $P(X \leq x_o) = P(Z \leq z_o)$

Normal Distribution : Solving problems

- **Complement Rule**:

$$P(Z \leq k) = 1 - P(Z \geq k)$$

  for some value k
- Alternatively $P(Z \geq k) = 1 - P(Z \leq k)$
- **Symmetry Rule**:

$$P(Z \leq -k) = P(Z \geq k)$$

  for some value k
- Alternatively $P(Z \geq -k) = P(Z \leq k)$

Normal Distribution : Solving problems

- **Intervals**:

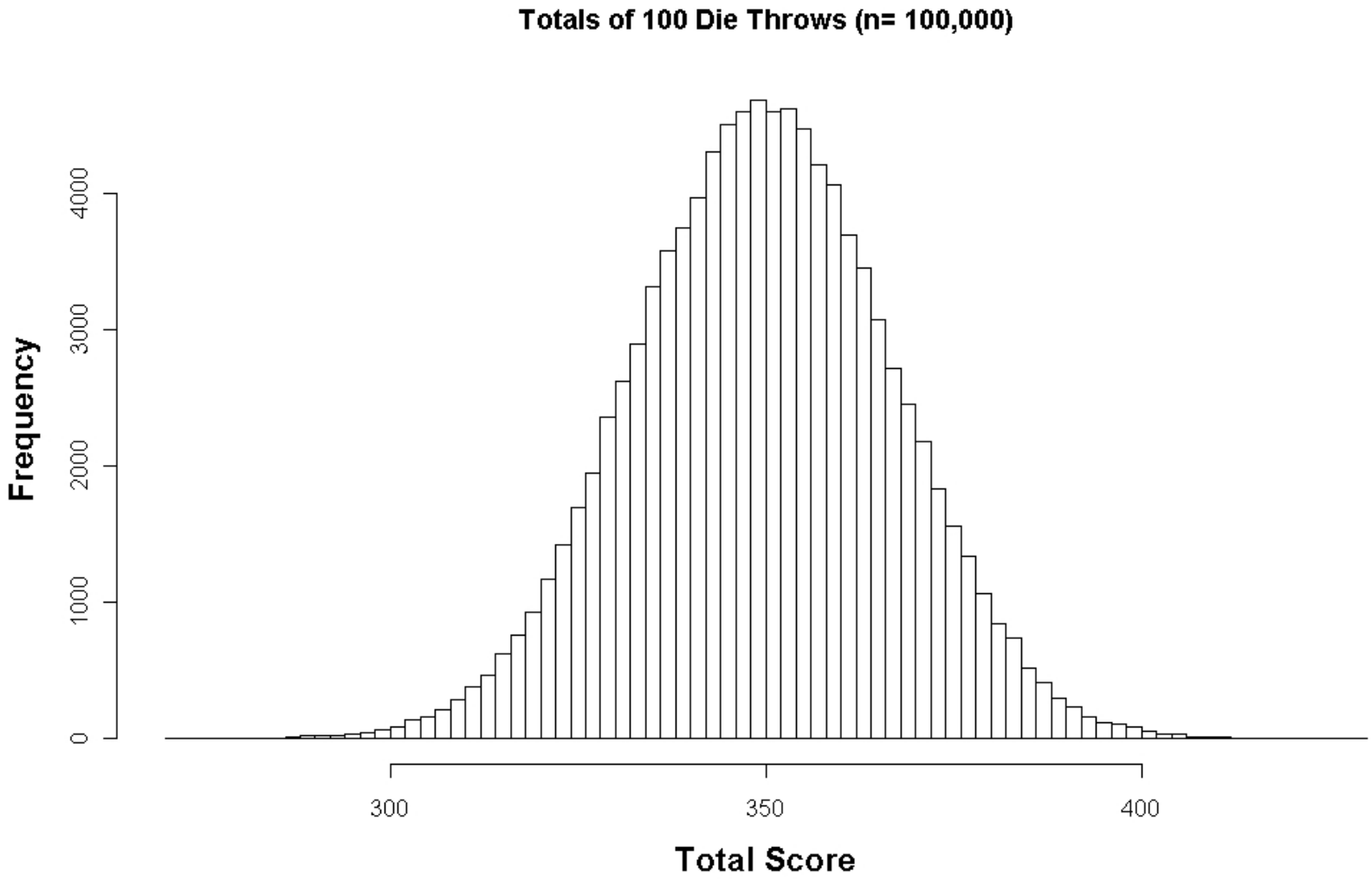
$$P(L \leq Z \leq U) = 1 - [P(Z \leq L) + P(Z \geq U)]$$

  where L and U are the lower and upper bounds of an interval.
- Probability of having a value too low for the interval : $P(Z \leq L)$
- Probability of having a value too high for the interval : $P(Z \geq U)$

Normal Distribution: Simulation Study

- Recall the experiment whereby a die was rolled 100 times, and the sum of the 100 values was recorded.
- This experiment was repeated a very large number of times (e.g. 100,000 times) in a simulation study.
- A histogram was drawn to depict the distribution of outcomes of this experiment.

Normal Distribution: Simulation Study



Normal Distribution: Simulation Study

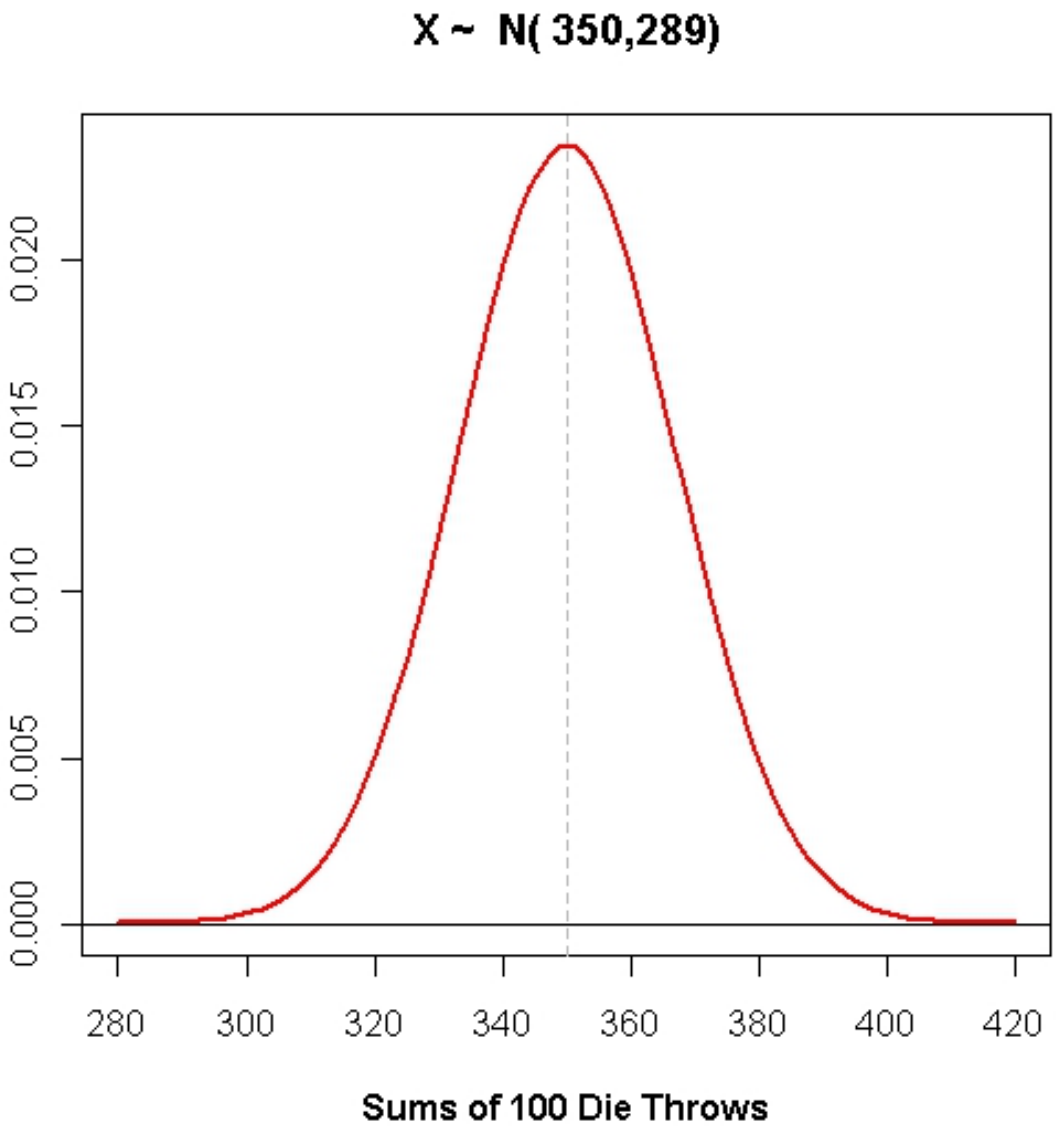
Recall some observations made about the results of the simulation study, made in a previous lecture.

- Approximately 68.7% of the values in the simulation study are between 332 and 367.
- Approximately 95% of the values are between 316 and 383.
- 2.5% of the values output are less than 316.
- 2.5% of the values study output are greater than 383.
- 175 values are greater than or equal to 400, whereas 198 values are less than or equal to 300.
- Results such as these are unusual, but they are not impossible.

Normal Distribution: Simulation Study

- Suppose we can ***approximate*** the summation of the die-throws using the normal distribution.
- The normal mean is necessarily $\mu = 350$.
- The normal standard deviation is approximately 17. (68% of values between 350 ± 17).
- Using the normal distribution, lets estimate the proportion of values greater than 383.

Normal Distribution: Simulation Study



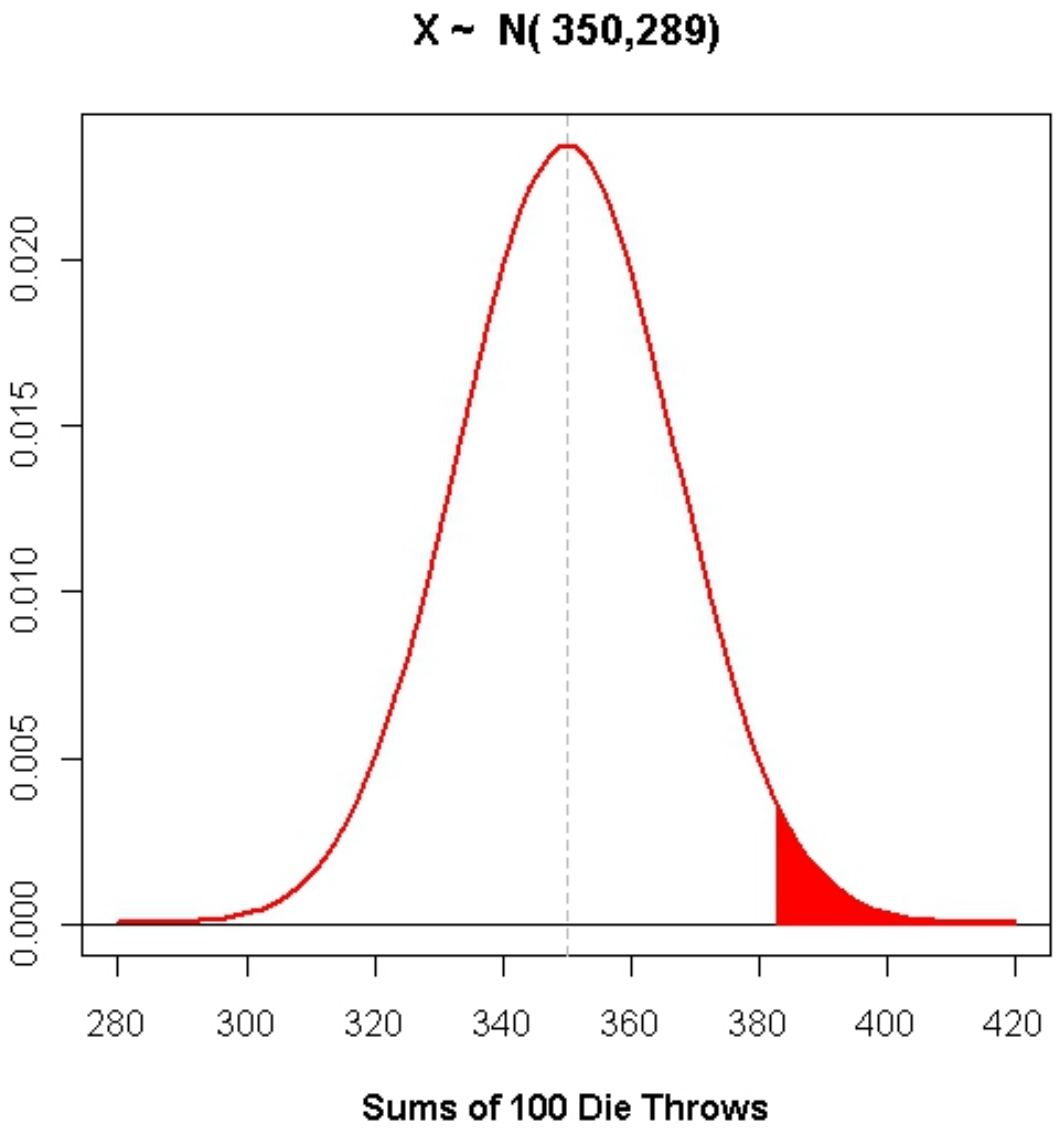
Normal Distribution: Simulation Study

- X is the normal random variable that approximates the sum of values from 100 throws of a die.
- Find $P(X \leq 383)$
- First use the standardization formula to find the Z-score.

$$z_o = \frac{383 - 350}{17} = \frac{33}{17} = 1.94$$

- Use the tables to compute $P(Z \geq 1.94)$ (Answer: 0.0262)
- Because $P(Z \geq 1.94) = 0.0262$, we can say $P(X \geq 383) = 0.0262$
- This is close to the proportion of observed values, which was 2.5%.
- Remark : The standard deviation of 17 was an estimate. The actual standard deviation should 17.12.

Normal Distribution: Simulation Study



Working Backwards

- Suppose we wish to find a value (lets call it A) from the normal distribution, such that a certain proportion of values is greater than A (e.g. 10%)
- Find A such that $P(X \geq A) = 0.10$. (with $\mu = 350$ and $\sigma = 17$)
- In general, our first step is to use the standardization equation to find the corresponding Z-score z_A.
- Because we don't know what value A has, we can't use this approach.
- However, we can say the following

$$P(X \geq A) = P(Z \geq z_A) = 0.10$$

- From the tables, we can approximate a value for z_A, by finding the closest probability value, and determining the corresponding Z-score.

Find z_A such that $P(Z \geq z_a) = 0.10$

- The closest probability value in the tables is 0.1003.
- The Z-score that corresponds to 0.1003 is 1.28.
- (Row : 1.2 , Column : 0.08)
- Therefore $z_A \approx 1.28$

	...	...	0.006	0.07	0.08	0.09
...	...	...	...	...	...	...
1.0	...	...	0.1446	0.1423	0.1401	0.1379
1.1	...	...	0.1230	0.1210	0.1190	0.1170
1.2	...	...	0.1038	0.1020	0.1003	0.0985
1.3	...	...	0.0869	0.0853	0.0838	0.0823
...	...	...	...	...	...	...

Working Backwards

- We can now use the standardization formula.
- We have only one unknown in the formula: A.

$$1.28 = \frac{A - 350}{17}$$

- Re-arranging (multiply both sides by 17):
  $21.76 = A - 350$
- Re-arranging (add 350 to both sides):
  $A = 371.76$
- $P(X \geq 371.76) \approx 0.10$
- (Remark: for sums of die-throws, round it to nearest value)

Working Backwards: Another Example

- Find B such that $P(X \geq B) = 0.90$. (with $\mu = 350$ and $\sigma = 17$)
- Necessarily $P(X \leq B) = 0.10$
- Find some value Z_B such that $P(Z \leq z_B) = 0.10$
- z_B could be negative.
- Use the symmetry rule $P(Z \leq z_B) = P(Z \geq -z_B)$
- $-z_B$ could be positive.
- Based on last example $-z_B = 1.28$. Therefore $z_B = -1.28$

Working Backwards

- Again ,we can now use the standardization formula
- We have only one unknown in the formula: B.

$$-1.28 = \frac{B - 350}{17}$$

- Re-arranging (multiply both sides by 17):
  $-21.76 = B - 350$
- Re-arranging (add 350 to both sides):
  $x_o = 350 - 21.76 = 328.24$
- $P(X \leq 328.24) \approx 0.10$

MA4413 Autumn 2008 paper

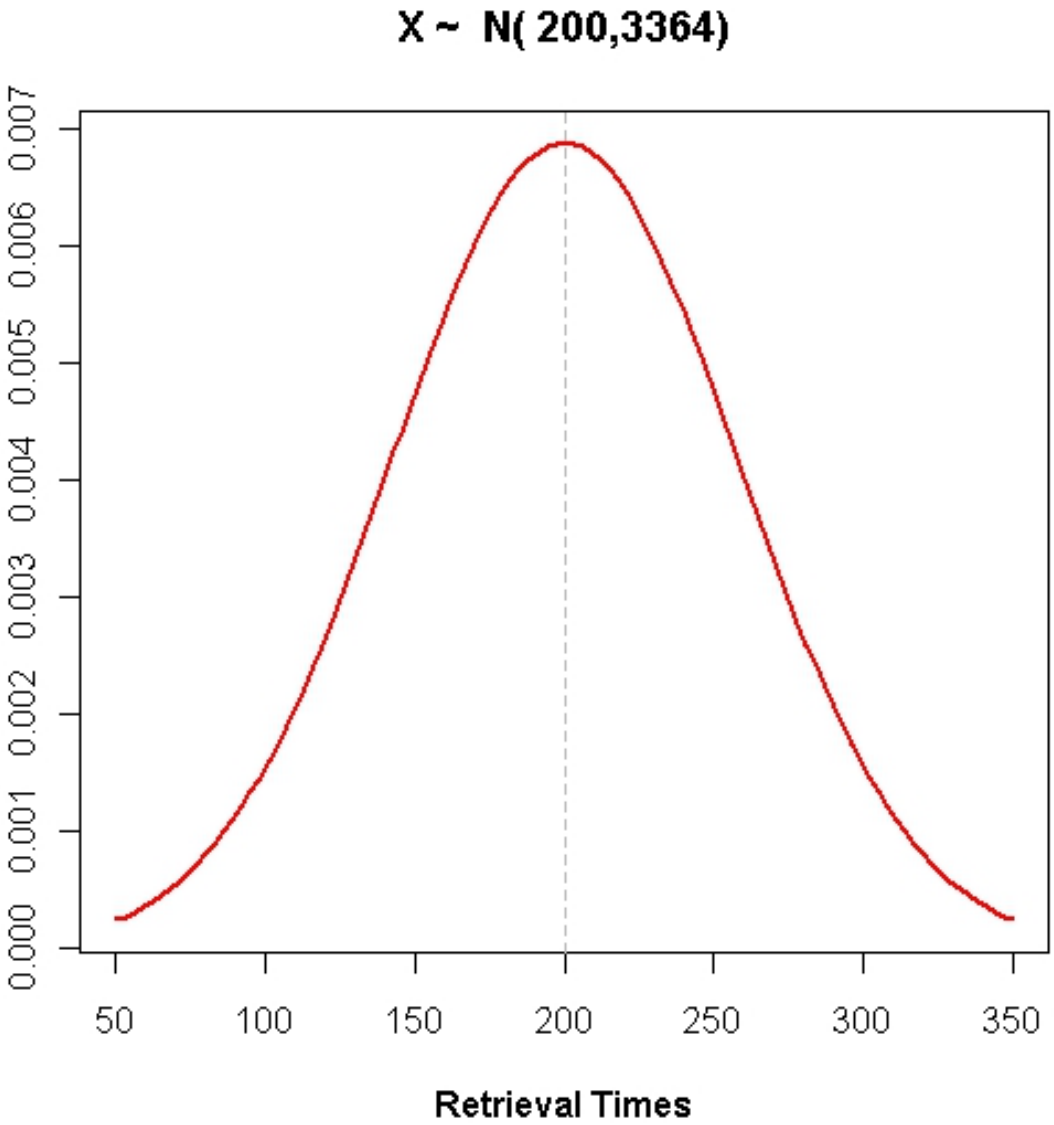
A model of an on-line computer system gives a mean times to retrieve a record from a direct access storage system device of 200 milliseconds, with a standard deviation of 58 milliseconds. If it can assumed that the retrieval times are normally distributed:

(i) What proportion of retrieval times will be greater than 75 milliseconds?

(ii) What proportion of retrieval times will be between 150 and 250 milliseconds?

(iii) What is the retrieval time below which 10% of retrieval times will be?

Normal Distribution



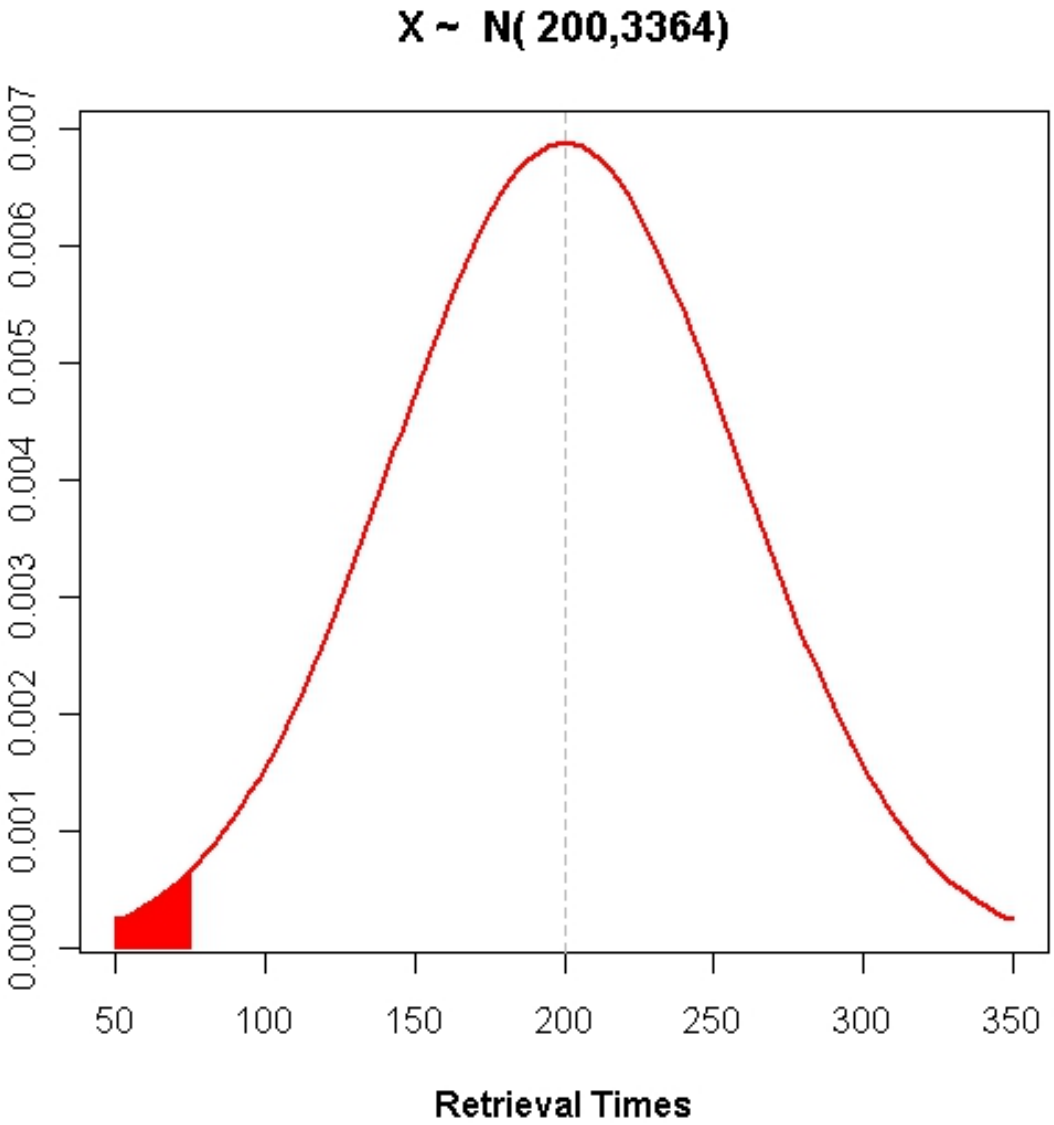
MA4413 Autumn 2008 paper (part 1)

What proportion of retrieval times will be greater than 75 milliseconds?

- Let X be the retrieval times, with $X \sim \mathrm{N}(200, 58^2)$.
- The first question asks us to find $P(X \geq 75)$.
- First compute the z score.

$$z_o = \frac{x_o - \mu}{\sigma} = \frac{75 - 200}{58} = -2.15$$

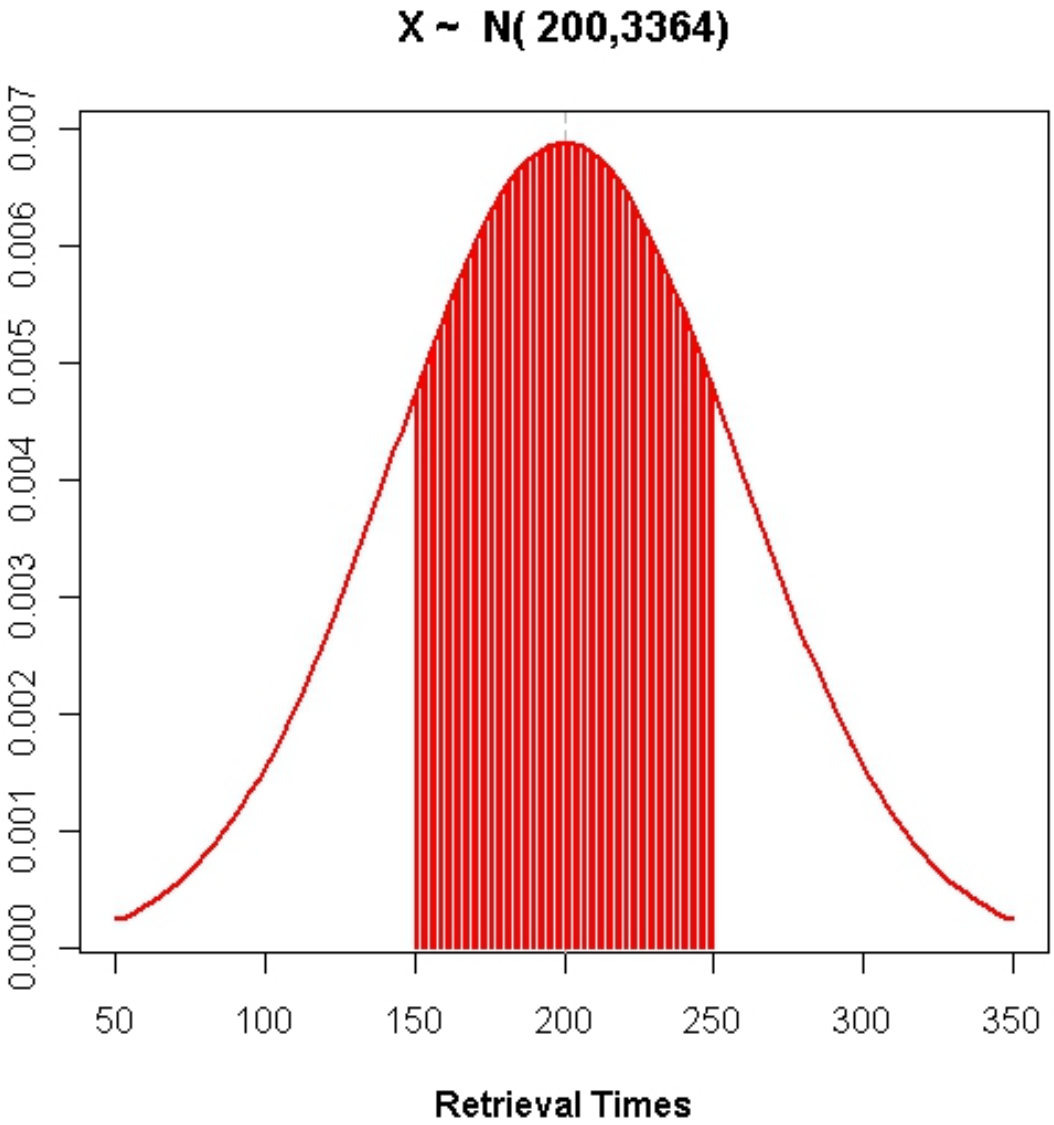
Normal Distribution



MA4413 Autumn 2008 paper (part 1)

- We can say
$$P(X \geq 75) = P(Z \geq -2.15)$$
- Using symmetry rule and complement rule
$$P(Z \geq -2.15) = P(Z \leq 2.15) = 1 - P(Z \geq 2.15)$$
- From tables $P(Z \geq 2.15) = 0.0158$
- Therefore $P(Z \leq 2.15) = 0.9842$
- Furthermore $P(X \geq 75) = \mathbf{0.9842}$ [Answer].

Normal Distribution



MA4413 Autumn 2008 paper (part 2)

- What proportion of retrieval times will be between 150 and 250 milliseconds?
- Find $P(150 \leq X \leq 250)$
- Use the 'Too Low / Too High ' approach.
- Too low $P(X \leq 150)$
- Too high $P(X \geq 250)$
- Find the z-scores for each.

$$z_{150} = \frac{150 - 200}{58} = -0.86$$

$$z_{250} = \frac{250 - 200}{58} = 0.86$$

MA4413 Autumn 2008 paper (part 2)

- We can now say

$$1. P(X \leq 150) = P(Z \leq -0.86)$$

$$2. P(X \geq 250) = P(Z \geq 0.86)$$

- By symmetry rule, $P(Z \leq -0.86) = P(Z \geq 0.86)$

$$P(X \leq 150) = P(X \geq 250)$$

- Let's compute $P(X \geq 250)$. Using tables

$$P(X \geq 250) = P(Z \geq 0.86) = 0.1949$$

MA4413 Autumn 2008 paper (part 2)

- Too high: $P(X \geq 250) = 0.1949$
- Too low: $P(X \leq 150) = 0.1949$
- Probability of being inside interval:

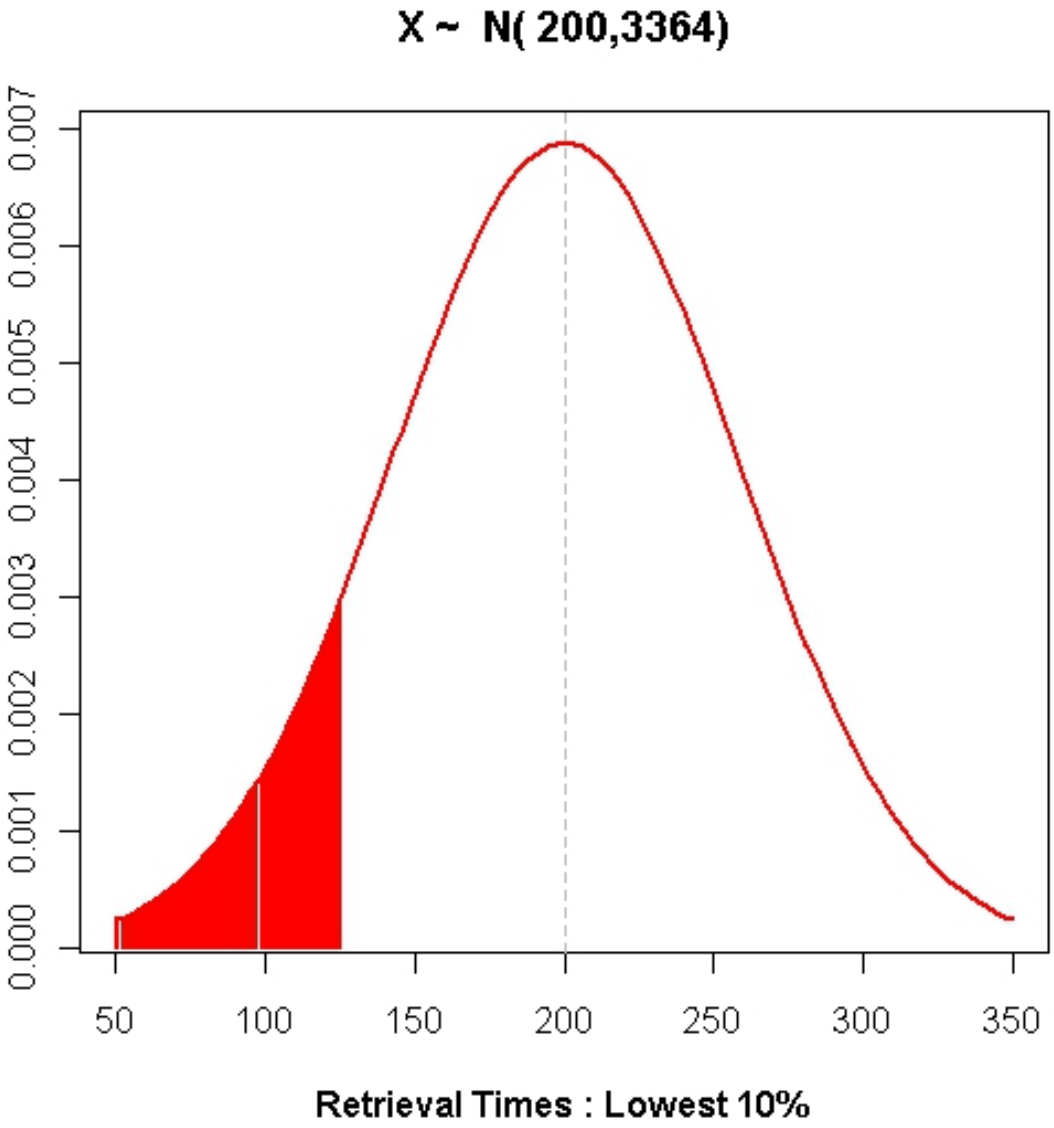
$$P(150 \leq X \leq 250) = 1 - [P(X \leq 150) + P(X \geq 250)]$$

- $P(150 \leq X \leq 250) = 1 - [0.1949 + 0.1949] = \mathbf{0.6102}$

MA4413 Autumn 2008 paper (part 3)

- What is the retrieval time below which 10% of retrieval times will be?
- Find A such that $P(X \leq A) = 0.10$.
- What z-score would correspond to A? Lets call it z_A.
- $P(Z \leq z_A) = 0.10$
- Remark: z_A could be negative.
- Using symmetry $P(Z \geq -z_A) = 0.10$
- Remark: $-z_A$ could be positive.

Normal Distribution



MA4413 Autumn 2008 paper (part 3)

- Use the Murdoch Barnes tables to get an approximate value for $-z_A$.
- The nearest value we can get is 1.28. ($P(Z \geq 1.28) = 0.1003$).
- If $-z_A = 1.28$, then $z_A = -1.28$
- We can now say

$$P(X \leq A) = P(Z \leq -1.28)$$

MA4413 Autumn 2008 paper (part 3)

- Necessarily A and Z_A are related by the standardization formula
- Recall that $\mu = 200$ and $\sigma = 58$.

$$-1.28 = \frac{A - 200}{58}$$

- Re-arranging (multiply both sides by 58)

$$-74.24 = A - 200$$

- Re-arranging again (Add 200 to both sides)

$$125.76 = A$$

MA4413 Autumn 2008 paper (part 3)

- Now we know the retrieval time below which 10% of retrieval times will be.
- $P(X \leq 125.76) = 0.10$ [Answer].