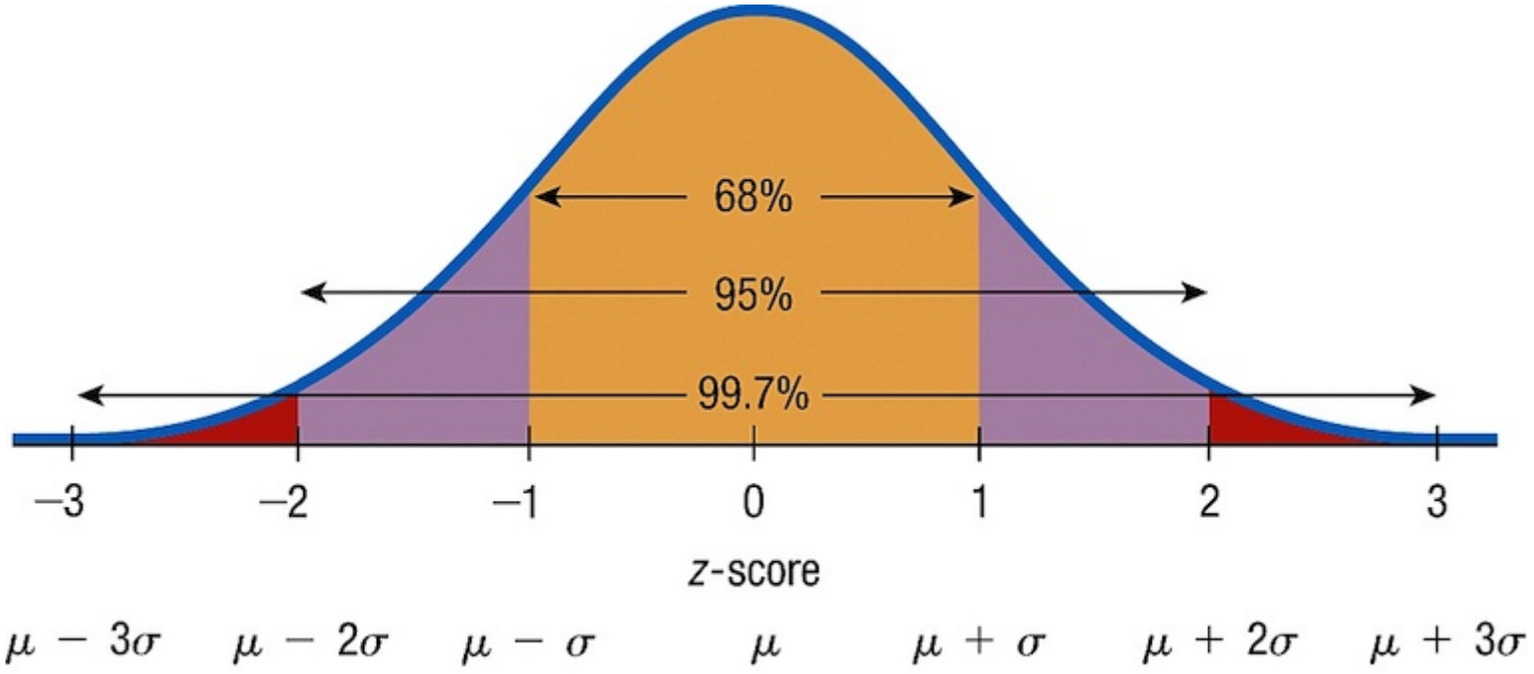
Lecture 7: Normal model

Criminology 1200

Prof Maria Cuellar

University of Pennslyvania

Normal model



Normal model

- The normal model is a probability model that is used to analyze data that conforms to a normal or Gaussian distribution. The normal distribution is also known as the Gaussian distribution. It is the most important and frequently used distribution in statistics.
- The normal model has two parameters: the population mean, μ, and the population standard deviation, σ. It is often written as $N(\mu, \sigma)$.
- The normal model is appropriate for *symmetric* and *unimodal* distributions. It is used to analyze data when there is an equally likely chance of being above or below the mean for continuous data whose histogram fits a bell curve.
- If distribution looks very different say in a histogram, the normal model should not be used.

What do we learn from the normal model?

- Data near the mean are more frequent in occurrence than data far from the mean.
- The mean helps to determine the line of symmetry of a graph.
- The standard deviation helps to know how far the data are spread out.
- Most values cluster around a central region and taper off as they go further away from the center

Standard Normal Model

If we model data with a Normal model and standardize them using the corresponding μ and σ, we still call the standardized value a z-score, and we write

$z = \frac{y-\mu}{\sigma}$.

Changing the center and spread of a variable is equivalent to changing its units. Indeed, the only part of the data's context changed by standardizing is the units.

We can think of the Normal in z-scores (numbers of standard deviations above or below the mean) or show the actual units.

Usually it's easier to standardize data using the mean and standard deviation first. Then we need only the model N(0, 1) with mean 0 and standard deviation 1. This Normal model is called the standard Normal model (or the standard Normal distribution).

What are these values? 68, 95, 99.7

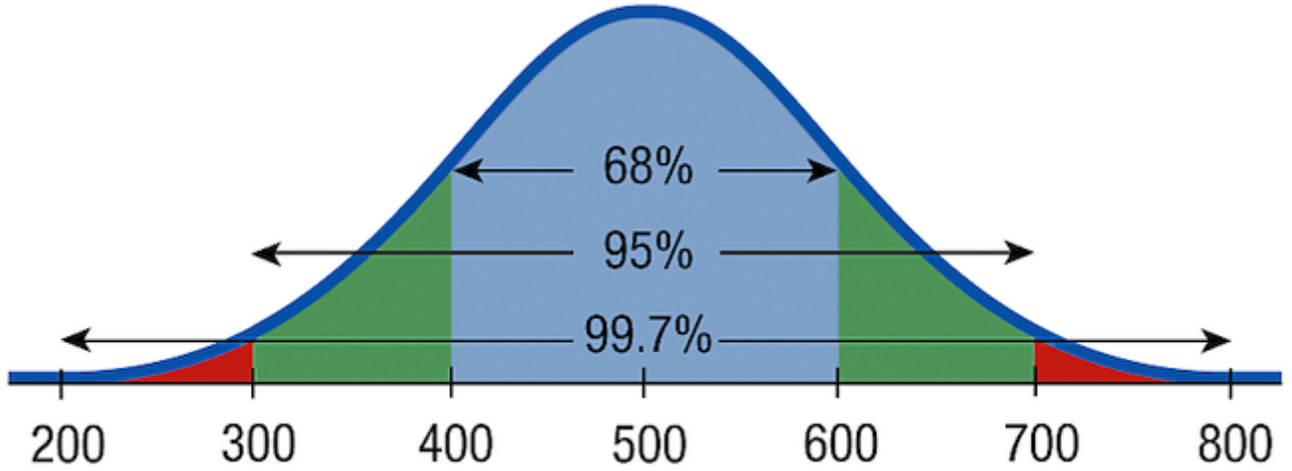
These magic values come from the Normal model. 68% means that that 68% of the data is under the curve within 2 standard deviations from the mean.

The normal model, can give us corresponding values for any number of z scores.

For example, it tells us that fewer than 1 out of a million values have z-scores smaller than -5 or larger than 5. So, if someone tells you you're "one in a million," they might be talking about your z-score.

SAT score example

- Each part of the SAT Reasoning Test has a distribution that is roughly unimodal and symmetric and is designed to have an overall mean of about 500 and a standard deviation of 100 for all test takers.



- Note that the bounds of SAT scoring at 200 and 800 can be explained by the 68–95–99.7 Rule. Because 200 and 800 are three standard deviations from 500, it hardly pays to extend the scoring any farther on either side. We'd get more information only on 100%-99.7%=0.3% of students.

- Suppose you earned a 600 on one part of your SAT. Where do you stand among all students who took that test?

SAT score example

- You could calculate your z-score and find out that it's z=(600-500)/100=1. But what does that tell you about your percentile? You'll need the Normal model and the 68–95–99.7 Rule to answer that question.
- I will model SAT score with a model. My score of 600 is 1 standard deviation above the mean. That corresponds to one of the points of the 68–95–99.7 Rule.
- About 68% of those who took the test had scores that fell no more than 1 standard deviation from the mean, so 100%-68%=32% of all students had scores more than 1 standard deviation away.
- Only half of those were on the high side, so about 16% (half of 32%) of the test scores were better than mine. My score of 600 is higher than about 84% of all scores on this test.

Percentiles

What if the score does not fall on one of the points of the 68-95-99.7 Rule?

We can use R!

pnorm() gives the proportion under the curve smaller than the observation and 1-pnorm() larger than the observation.

For example, this is the area under the curve at value 600

```
pnorm(600, mean=500, sd=100)
```

```
## [1] 0.8413447
```

Magic points

We can use pnorm() to show the 68-95-99.7 points:

```
2*(pnorm(1, mean=0, sd=1)-.5)
```

```
## [1] 0.6826895
```

```
2*(pnorm(2, mean=0, sd=1)-.5)
```

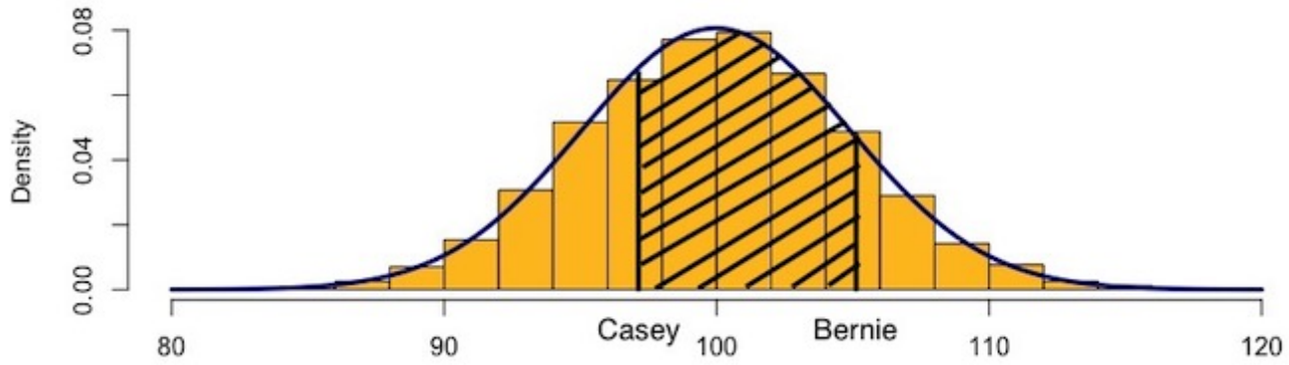
```
## [1] 0.9544997
```

```
2*(pnorm(3, mean=0, sd=1)-.5)
```

```
## [1] 0.9973002
```

Using pnorm

Find the area under the curve for these data, excluding the crossed out section.




```
pnorm(97.3, mean = 100, sd = 5) + 1 - pnorm(105.6, mean = 100, sd = 5)
```

```
## [1] 0.4259554
```

More on this here: http://rstudio-pubs-static.s3.amazonaws.com/383738_49fb6c7c59f74756aece1a8803fb828f.html