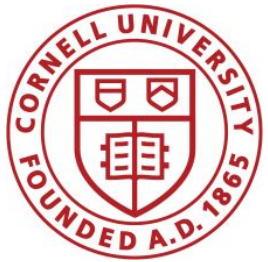


DSFA

Spring 2019

Lecture 21

The Normal Distribution

Announcements

- Homework 6: Due Thursday, April 25

Questions

- How can we quantify natural concepts like “center” and “variability”?
- Why do many of the empirical distributions that we generate come out bell shaped?
- How is sample size related to the accuracy of an estimate?

Standard Deviation (Review)

How Far from the Average?

- Standard deviation (SD) measures roughly how far the data are from their average
- SD = root mean square of deviations from average

  5 4 3 2 1
- SD has the same units as the data

Why Use the SD?

There are two main reasons.

- **The first reason:**

No matter what the shape of the distribution,
the bulk of the data are in the range “average ± a few SDs”

- **The second reason:**

Coming up later in this lecture ...

How Big are Most of the Values?

No matter what the shape of the distribution,
the bulk of the data are in the range “average ± a few SDs”

Chebyshev’s Inequality
No matter what the shape of the distribution,
the proportion of values in the range “average ± z SDs” is

at least $1 - 1/z^2$

Chebyshev's Bounds

Range	Proportion
average ± 2 SDs	at least 1 - 1/4 (75%)
average ± 3 SDs	at least 1 - 1/9 (88.888…%)
average ± 4 SDs	at least 1 - 1/16 (93.75%)
average ± 5 SDs	at least 1 - 1/25 (96%)

No matter what the distribution looks like

(Demo)

Standard Units

Standard Units

- How many SDs above average?
- ***z* = (value - mean)/SD**
  - Negative z: value below average
  - Positive z: value above average
  - z = 0: value equal to average
- When values are in standard units: average = 0, SD = 1
- Chebyshev: At least 96% of the values of *z* are between -5 and 5

(Demo)

Discussion Question

Find whole numbers that are close to:

(a) the average age

(b) the SD of the ages

(Demo)

Age in Years	Age in Standard Units
27	-0.0392546
33	0.992496
28	0.132704
23	-0.727088
25	-0.383171
33	0.992496
23	-0.727088
25	-0.383171
30	0.476621
27	-0.0392546

... (1164 rows omitted)

The SD and the Histogram

- Usually, it's not easy to estimate the SD by looking at a histogram.
- But if the histogram has a bell shape, then you can.

The SD and Bell-Shaped Curves

If a histogram is bell-shaped, then

- the average is at the center
- the SD is the distance between the average and the points of inflection on either side

The Normal Distribution

:6762761K7

che Bundesbank

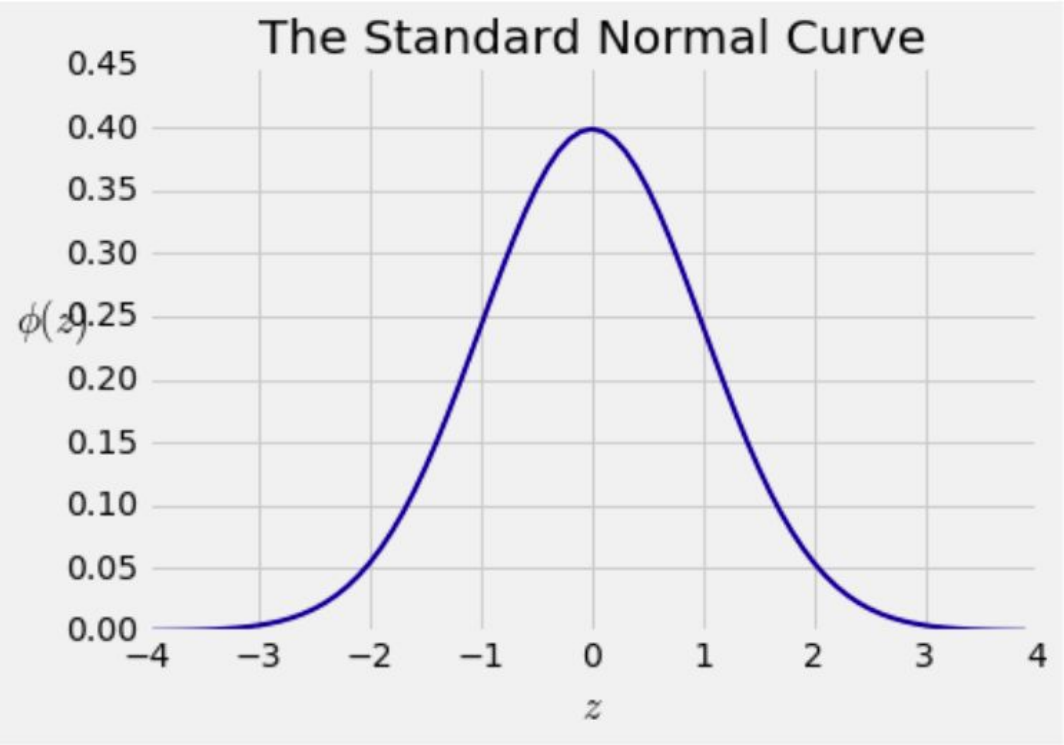
urt am Main
ıar 1989

The Standard Normal Curve

A beautiful formula that we won't use at all:

$$\phi(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}, \qquad -\infty < z < \infty$$

Bell Curve


The Standard Normal Curve
0.45
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
$\phi(z)$
−4
−3
−2
−1
0
1
2
3
4
z

Normal Proportions

How Big are Most of the Values?

No matter what the shape of the distribution,
the bulk of the data are in the range “average ± a few SDs”

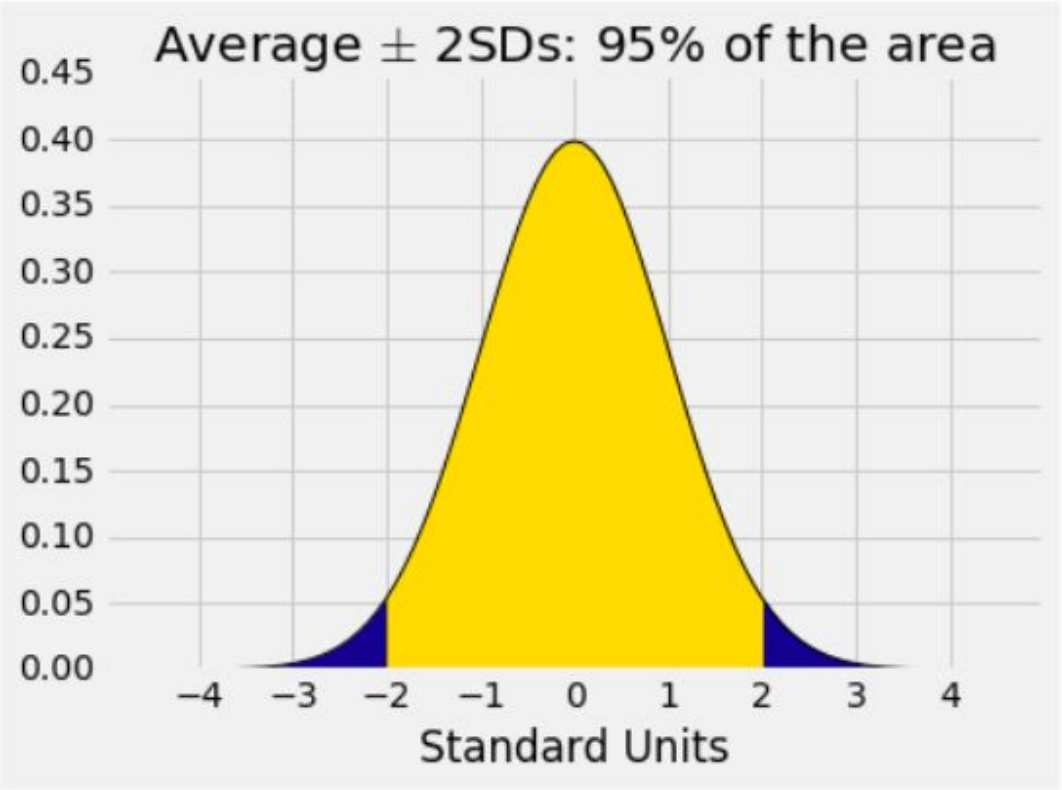
If a histogram is bell-shaped, then

- Almost all of the data are in the range “average ± 3 SDs”

Bounds and Normal Approximations

Percent in Range	All Distributions	Normal Distribution
average $\pm$ 1 SD	at least 0%	about 68%
average $\pm$ 2 SDs	at least 75%	about 95%
average $\pm$ 3 SDs	at least 88.888...%	about 99.73%

A “Central” Area



(Demo)

Central Limit Theorem

Second Reason for Using the SD

If the sample is

- large, and
- drawn at random with replacement,

Then, *regardless of the distribution of the population,*

the probability distribution of the sample sum (or of the sample average) is roughly bell-shaped

(Demo)

Sample Averages

- The Central Limit Theorem describes how the normal distribution (a bell-shaped curve) arises in the context of random sampling.
- Most distributions we observed were not bell-shaped, but empirical distributions of sample averages were.
- We care about sample averages because they estimate population averages.

Distribution of the Sample Average

Why is There a Distribution?

- You have only one random sample, and it has only one average.
- But **the sample could have come out differently**.
- And then the sample average might have been different.
- So there are many possible sample averages.

Distribution of the Sample Average

- Imagine all possible random samples of the same size as yours. There are lots of them.
- Each of these samples has a mean.
- The **distribution of the sample average** is the distribution of the means of all the possible samples.

Shape of the Distribution

Central Limit Theorem

If the sample is

- large, and
- drawn at random with replacement,

Then, *regardless of the distribution of the population,*

the distribution of the sample sum
(or of the sample average) is roughly bell-shaped

(Demo)

Specifying the Distribution

Suppose the random sample is large.

- We have seen that the distribution of the sample average is roughly bell shaped.

- Important questions remain:
  - Where is the center of that bell curve?
  - How wide is that bell curve?

Center of the Distribution

The Population Average

The distribution of the sample average is roughly a bell curve centered at the population average.

Variability of the Sample Average

Why Is This Important?

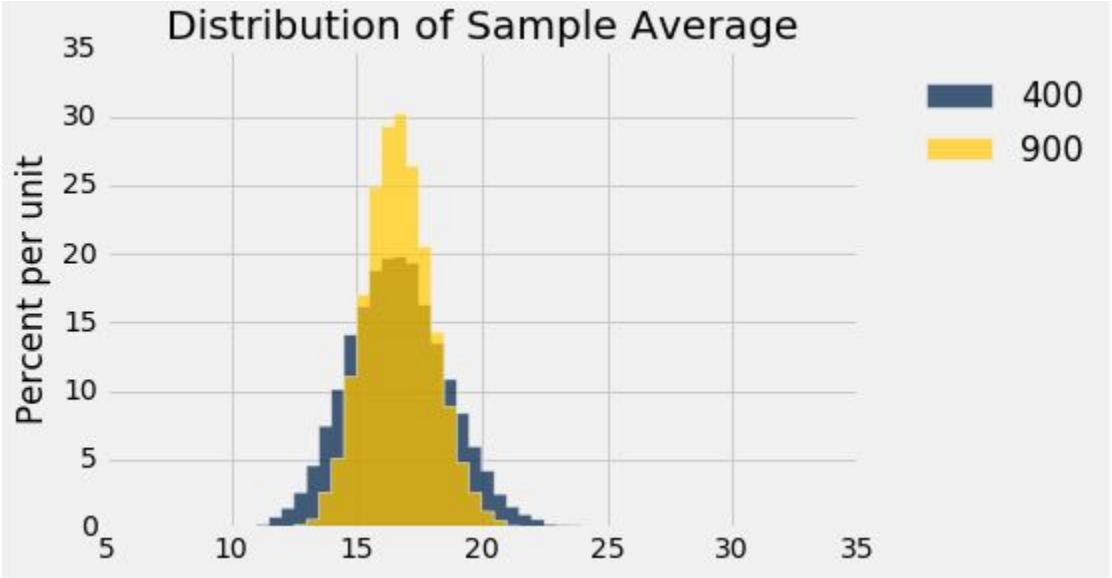
- Along with the center, the spread helps identify exactly which normal curve is the distribution of the sample average.
- The variability of the sample average helps us measure how accurate the sample average is as an estimate of the population average.
- If we want a specified level of accuracy, understanding the variability of the sample mean helps us work out how large our sample has to be.

(Demo)

Discussion Question

The gold histogram shows the distribution of _____ values, each of which is _____.

(a) 900
(b) 10,000
(c) a randomly sampled flight delay
(d) an average of flight delays



The Two Histograms

- The gold histogram shows the distribution of 10,000 values, each of which is an average of 900 randomly sampled flight delays.
- The blue histogram shows the distribution of 10,000 values, each of which is an average of 400 randomly sampled flight delays.
- Both are roughly bell shaped.
- The larger the sample size, the narrower the bell.

(Demo)

Variability of the Sample Average

- Fix a large sample size.
- Draw all possible random samples of that size.
- Compute the average of each sample.
- You'll end up with a lot of averages.
- The distribution of those is called the *distribution of the sample average.*
- It's roughly normal, centered at the population average.
- SD = (population SD) / $\sqrt{\text{sample size}}$

Discussion Question

A city has 200,000 households. The annual incomes of these households have an average of $65,000 and an SD of $45,000. The distribution of the incomes [pick one and explain]:

(a) is roughly normal because the number of households is large.
(b) is not close to normal.
(c) may be close to normal, or not; we can't tell from the information given.