Lecture 4:
Normal curve and inferential statistics

Ernesto F. L. Amaral

September 15, 2022
Introduction to Sociological Data Analysis (SOCI 600)

Source: Healey, Joseph F. 2015. ”Statistics: A Tool for Social Research.” Stamford: Cengage Learning. 10th edition. Chapters 5 (pp. 122–142), 6 (pp. 144–159), 7 (pp. 160–184).

TEXAS A&M UNIVERSITY.

Outline

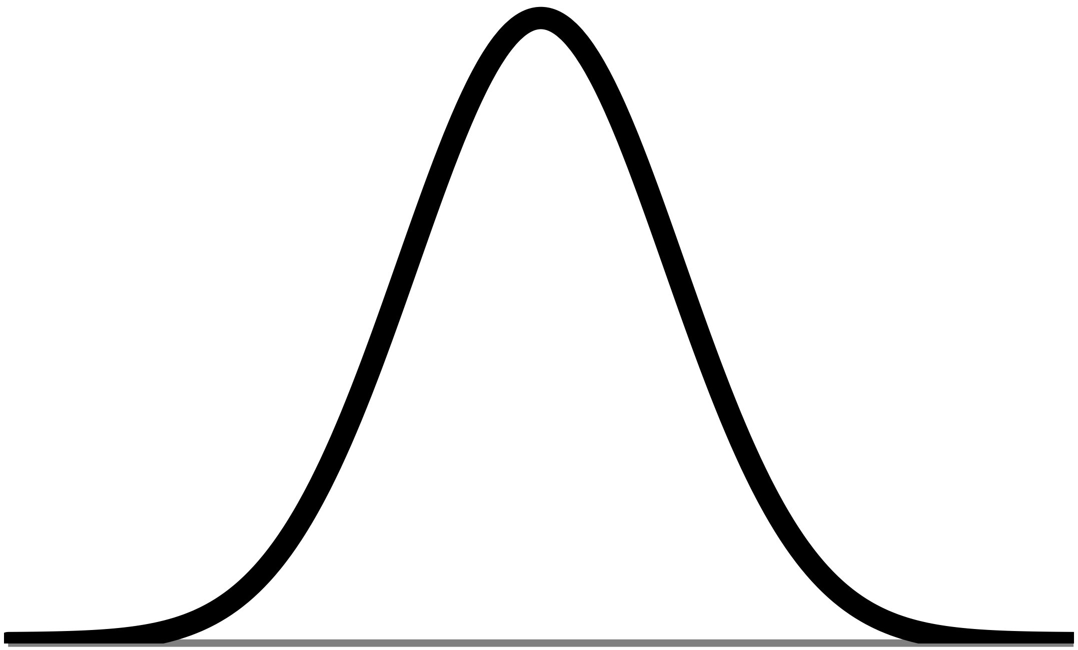
- The normal curve
- Inferential statistics
  - Sampling
  - The sampling distribution
- Estimation procedures

The normal curve

- Define and explain the concept of the normal curve
- Convert empirical scores to Z scores
- Use Z scores and the normal curve table (Appendix A) to find areas above, below, and between points on the curve
- Express areas under the curve in terms of probabilities

Properties of the normal curve

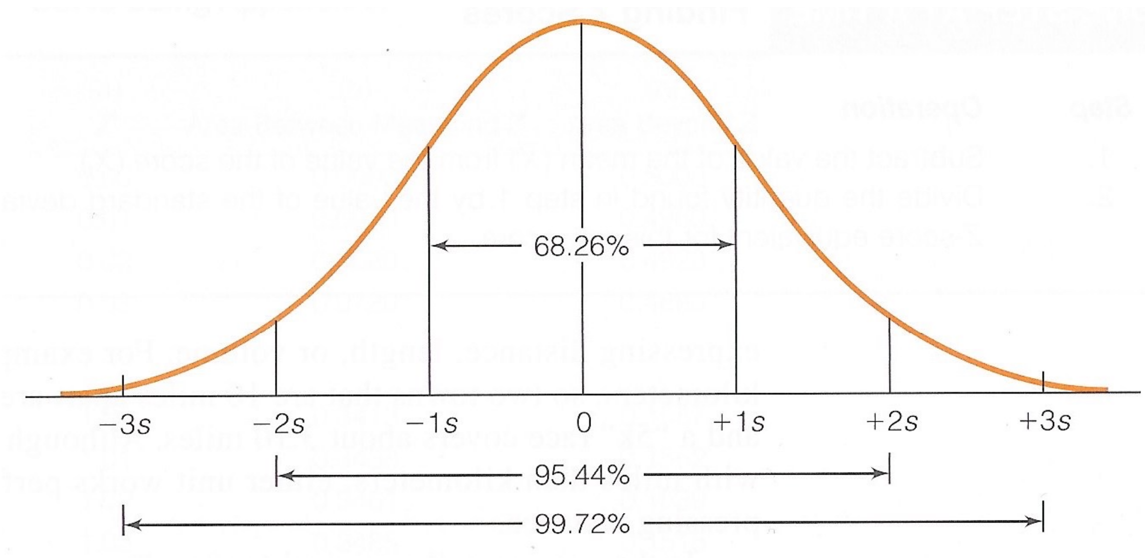
- Theoretical
- Bell-shaped
- Unimodal
- Smooth
- Symmetrical
- Unskewed



- Tails extend to infinity
- Mode, median, and mean are same value

Standard normal distribution

- Normal distribution with $\bar{X} = 0$ and $s = 1$
  - Distances on horizontal axis cut off the same area



- ±1s = 68.26%
- ±2s = 95.44%
- ±3s = 99.72%

- Between mean & 1s = 34.13%
- Between mean & 2s = 47.72%
- Between mean & 3s = 49.86%

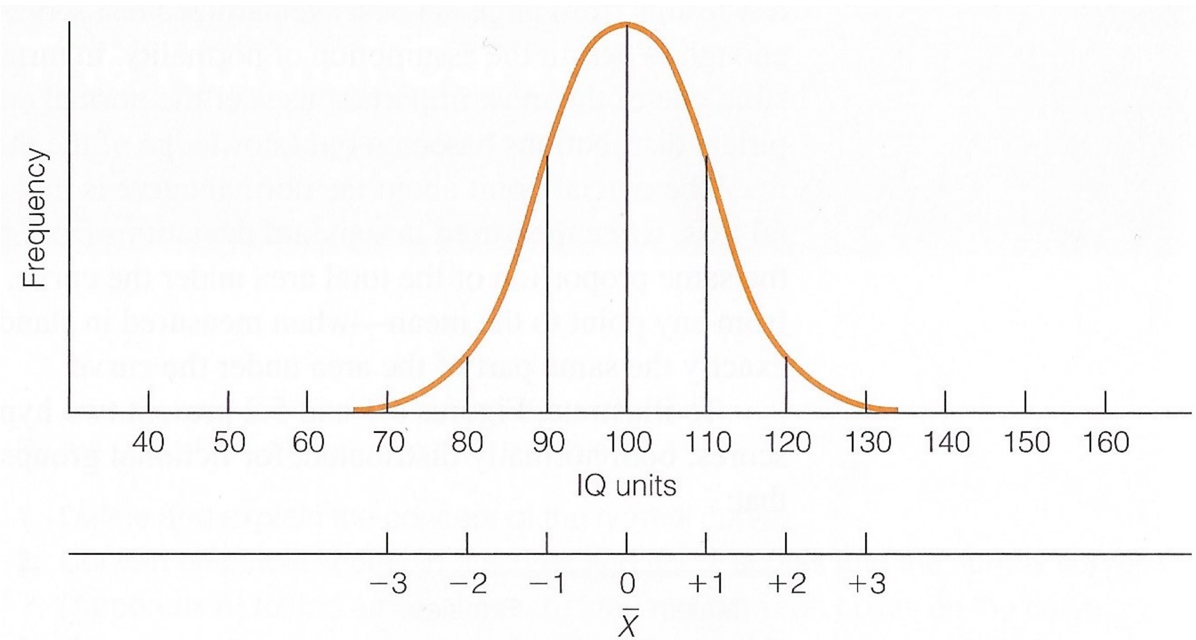
Source: Healey 2015, p.125.

IQ scores, females

$\bar{X}$ = 100

s = 10

N = 1000

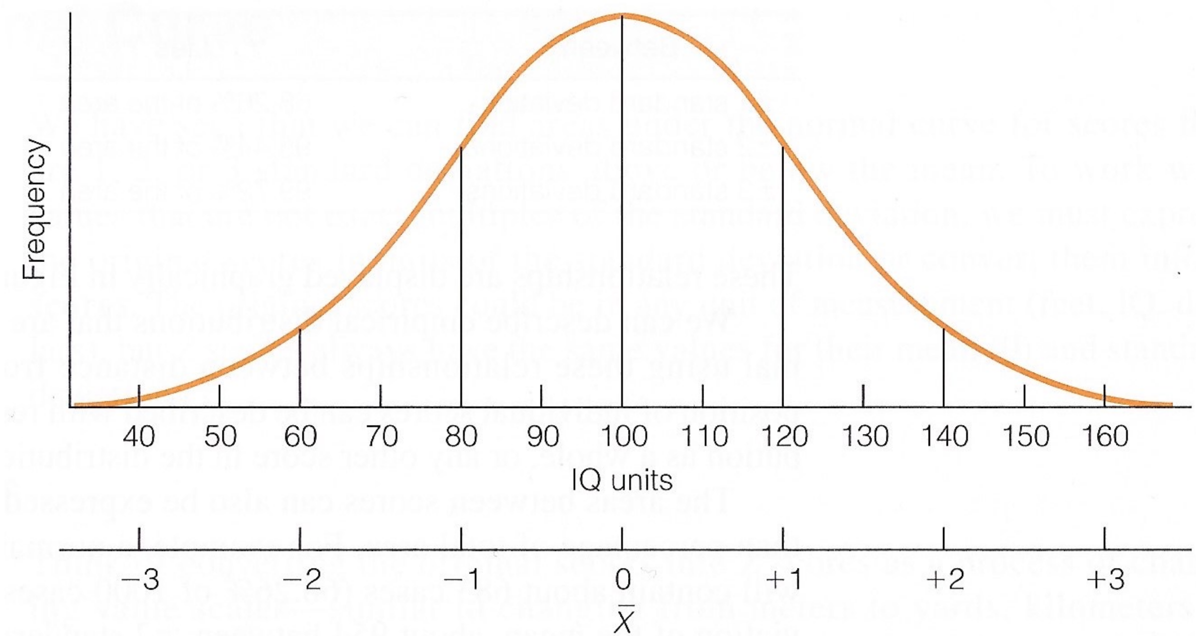


IQ scores, males

$\bar{X}$ = 100

s = 20

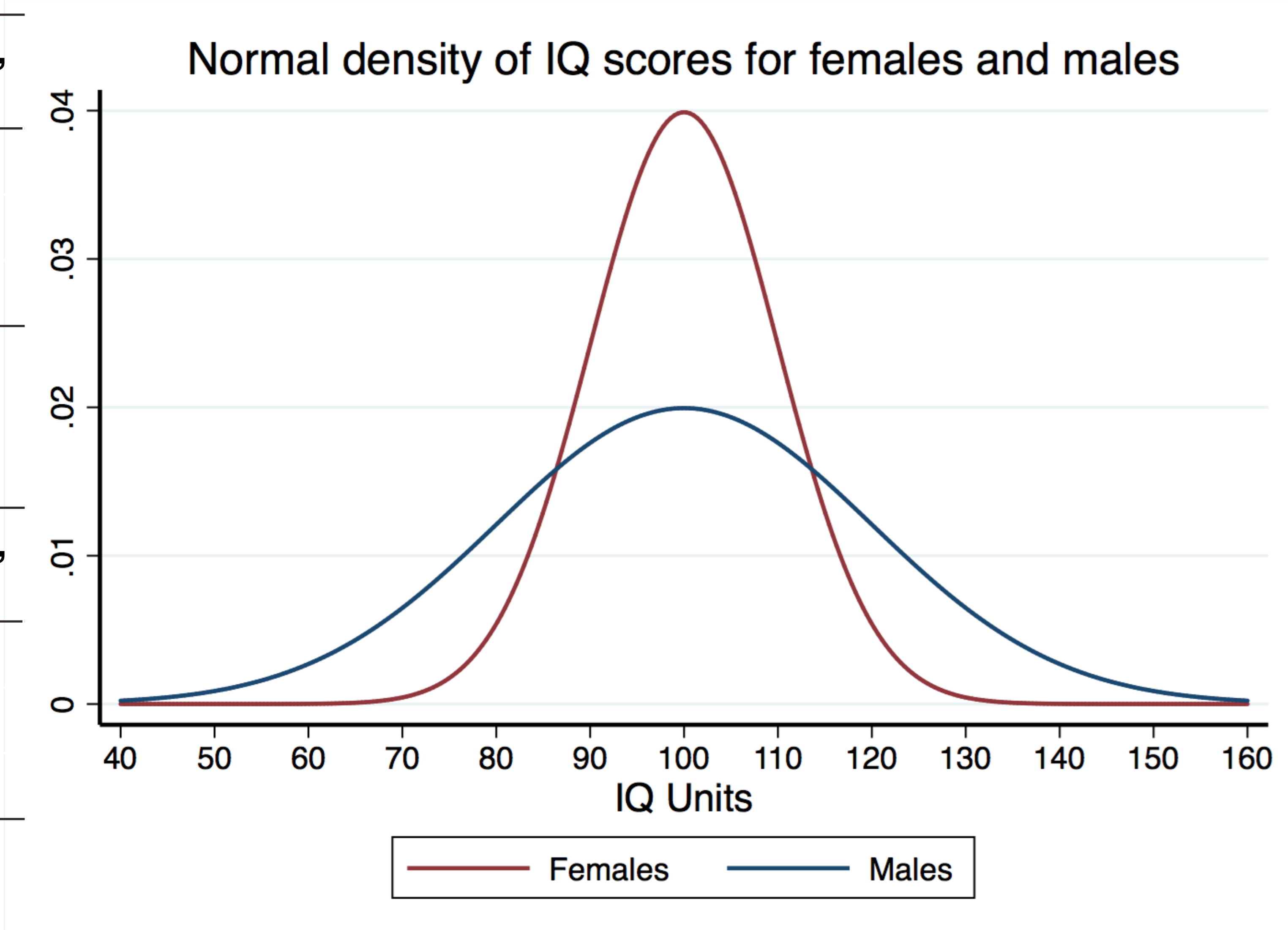
N = 1000



Source: Healey 2015, p.123–124.

IQ scores, females
$\bar{X}$ = 100
s = 10
N = 1000

IQ scores, males
$\bar{X}$ = 100
s = 20
N = 1000



Source: Healey 2015, p.123–124.

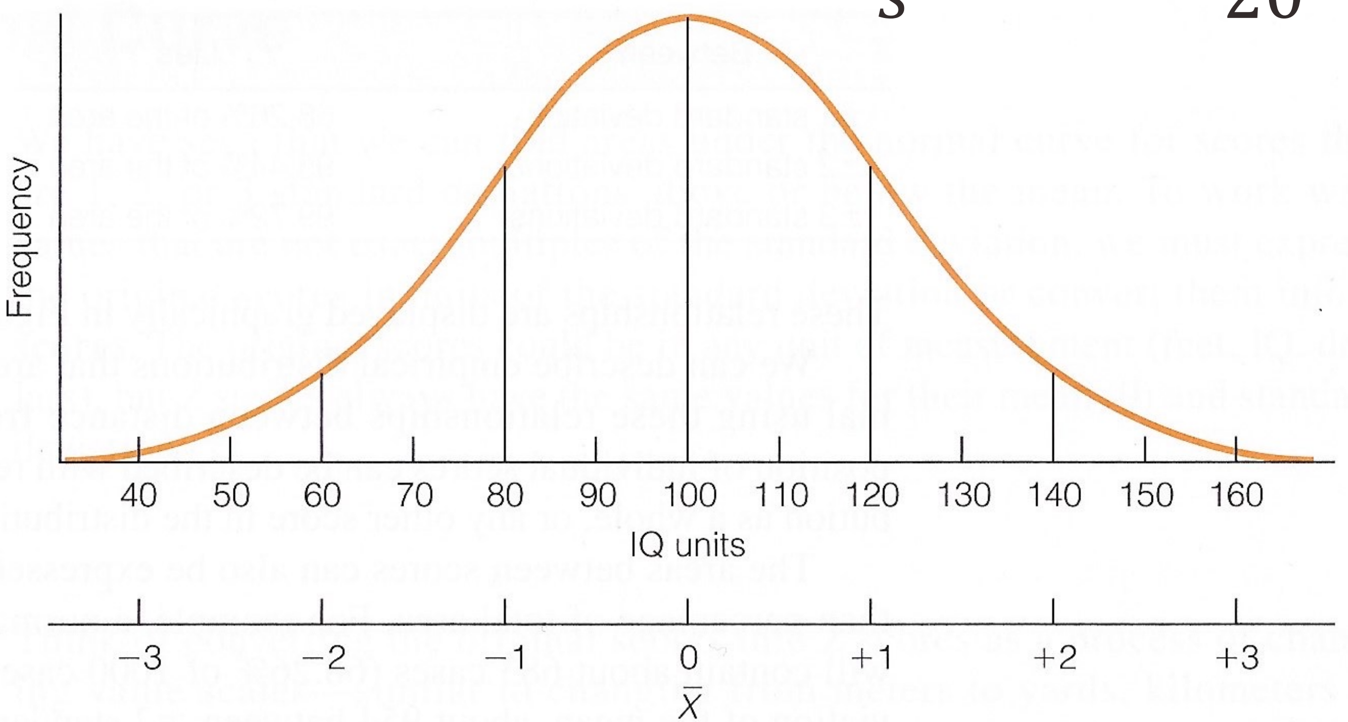
Z scores

- Z scores are scores that have been standardized to the theoretical normal curve
- Z scores represent how different a raw score is from the mean in standard deviation units
- To find areas, first compute Z scores
- The Z score formula changes a raw score to a standardized score

$$Z = \frac{X_i - \bar{X}}{s}$$

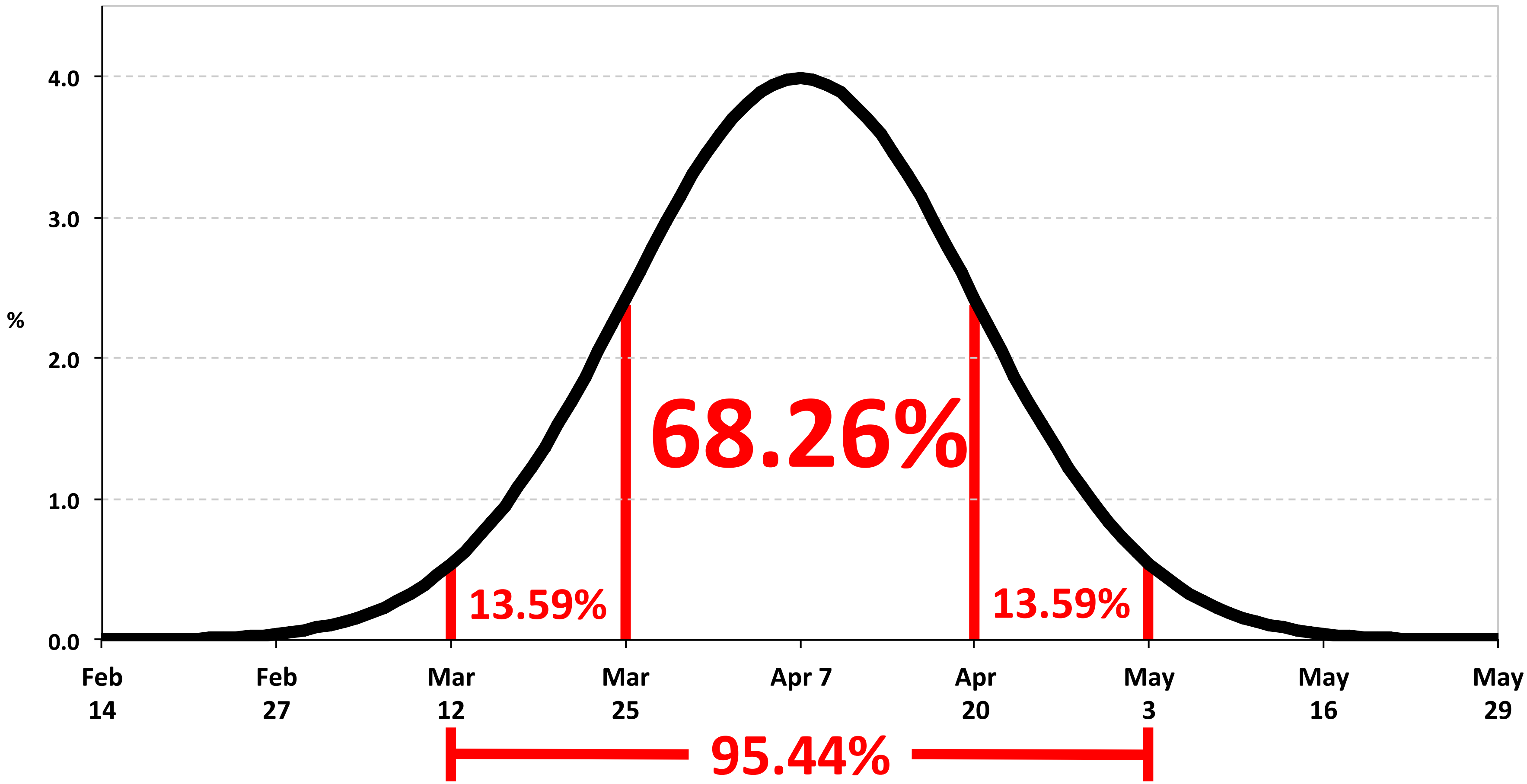
IQ for males

$$Z = \frac{X_i - \bar{X}}{s} = \frac{120 - 100}{20} = +1.00$$



- An IQ score of 120 falls one standard deviation above (to the right of) the mean

Estimated date of delivery



s = 13 days (based on Naegele's rule)

Area under the normal curve

- Compute the Z score
- Draw a picture of the normal curve and shade in the area in which you are interested
- Find your Z score in Column A...

FIGURE A.1 Area Between Mean and *Z*

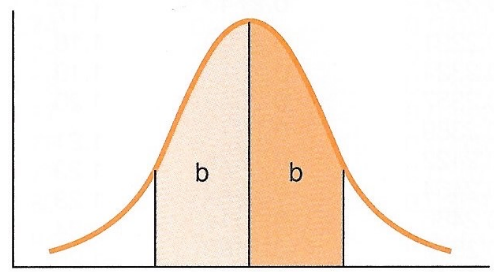
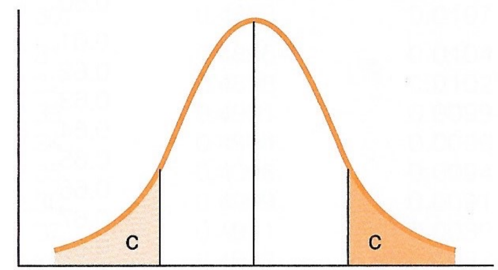


FIGURE A.2 Area Beyond *Z*



(a) *Z*	(b) Area Between Mean and *Z*	(c) Area Beyond *Z*
0.00	0.0000	0.5000
0.01	0.0040	0.4960
0.02	0.0080	0.4920
0.03	0.0120	0.4880
0.04	0.0160	0.4840
0.05	0.0199	0.4801
0.06	0.0239	0.4761
0.07	0.0279	0.4721
0.08	0.0319	0.4681
0.09	0.0359	0.4641
0.10	0.0398	0.4602
0.11	0.0438	0.4562
0.12	0.0478	0.4522
0.13	0.0517	0.4483
0.14	0.0557	0.4443
0.15	0.0596	0.4404
0.16	0.0636	0.4364
0.17	0.0675	0.4325
0.18	0.0714	0.4286
0.19	0.0753	0.4247
0.20	0.0793	0.4207

(a) *Z*	(b) Area Between Mean and *Z*	(c) Area Beyond *Z*
0.21	0.0832	0.4168
0.22	0.0871	0.4129
0.23	0.0910	0.4090
0.24	0.0948	0.4052
0.25	0.0987	0.4013
0.26	0.1026	0.3974
0.27	0.1064	0.3936
0.28	0.1103	0.3897
0.29	0.1141	0.3859
0.30	0.1179	0.3821
0.31	0.1217	0.3783
0.32	0.1255	0.3745
0.33	0.1293	0.3707
0.34	0.1331	0.3669
0.35	0.1368	0.3632
0.36	0.1406	0.3594
0.37	0.1443	0.3557
0.38	0.1480	0.3520
0.39	0.1517	0.3483
0.40	0.1554	0.3446
...	...	...

Source: Healey 2015, Appendix A, p.443.

Positive score

- Find your Z score in Column A
- To find area below a positive score
  - Add column b area to 0.50
- To find area above a positive score
  - Look in column c

FIGURE A.1 Area Between Mean and *Z*

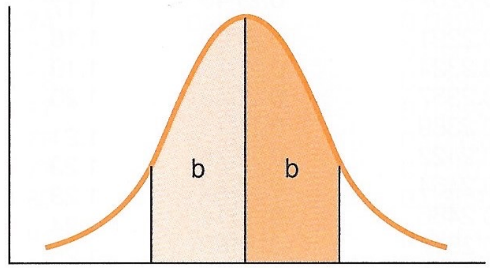
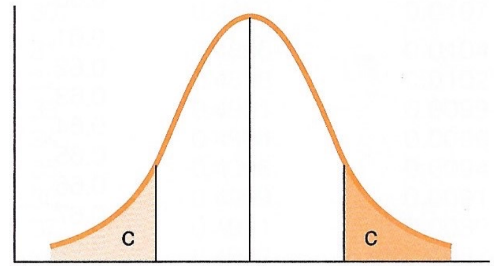


FIGURE A.2 Area Beyond *Z*



(a) *Z*	(b) Area Between Mean and *Z*	(c) Area Beyond *Z*
0.00	0.0000	0.5000
0.01	0.0040	0.4960
0.02	0.0080	0.4920
0.03	0.0120	0.4880
0.04	0.0160	0.4840
0.05	0.0199	0.4801
0.06	0.0239	0.4761
0.07	0.0279	0.4721
0.08	0.0319	0.4681
0.09	0.0359	0.4641
0.10	0.0398	0.4602
0.11	0.0438	0.4562
0.12	0.0478	0.4522
0.13	0.0517	0.4483
0.14	0.0557	0.4443
0.15	0.0596	0.4404
0.16	0.0636	0.4364
0.17	0.0675	0.4325
0.18	0.0714	0.4286
0.19	0.0753	0.4247
0.20	0.0793	0.4207

(a) *Z*	(b) Area Between Mean and *Z*	(c) Area Beyond *Z*
0.21	0.0832	0.4168
0.22	0.0871	0.4129
0.23	0.0910	0.4090
0.24	0.0948	0.4052
0.25	0.0987	0.4013
0.26	0.1026	0.3974
0.27	0.1064	0.3936
0.28	0.1103	0.3897
0.29	0.1141	0.3859
0.30	0.1179	0.3821
0.31	0.1217	0.3783
0.32	0.1255	0.3745
0.33	0.1293	0.3707
0.34	0.1331	0.3669
0.35	0.1368	0.3632
0.36	0.1406	0.3594
0.37	0.1443	0.3557
0.38	0.1480	0.3520
0.39	0.1517	0.3483
0.40	0.1554	0.3446
...	...	...

Source: Healey 2015, Appendix A, p.443.

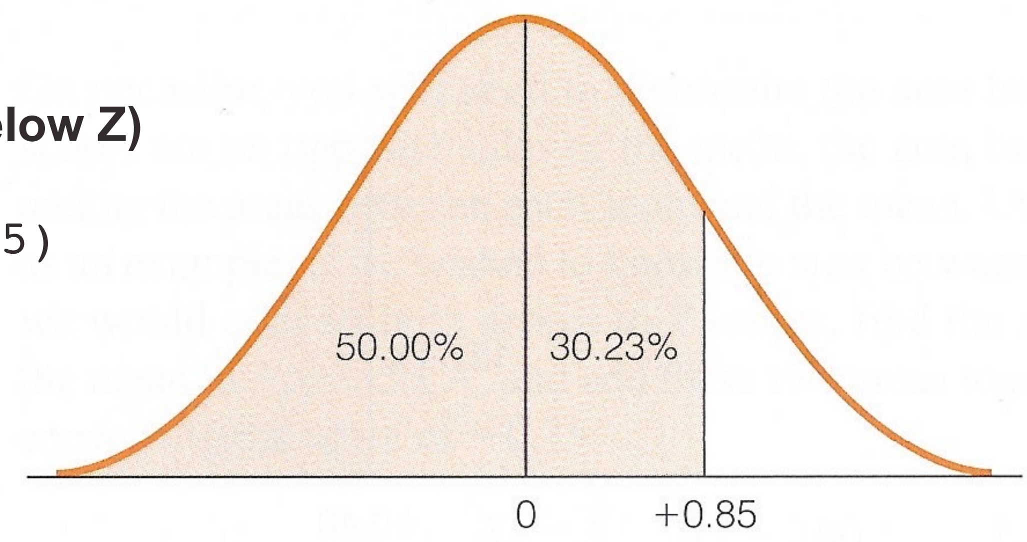
Area below Z = 0.85

- Finding the area below a positive Z score:
  - Z = +0.85
  - Area from column b = 0.3023
  - 0.50 + 0.3023 = 0.8023 or 80.23%

Command in Stata
(normal shows area below Z)

```
display normal(0.85)

.80233746
```



Source: Healey 2015, p.129.

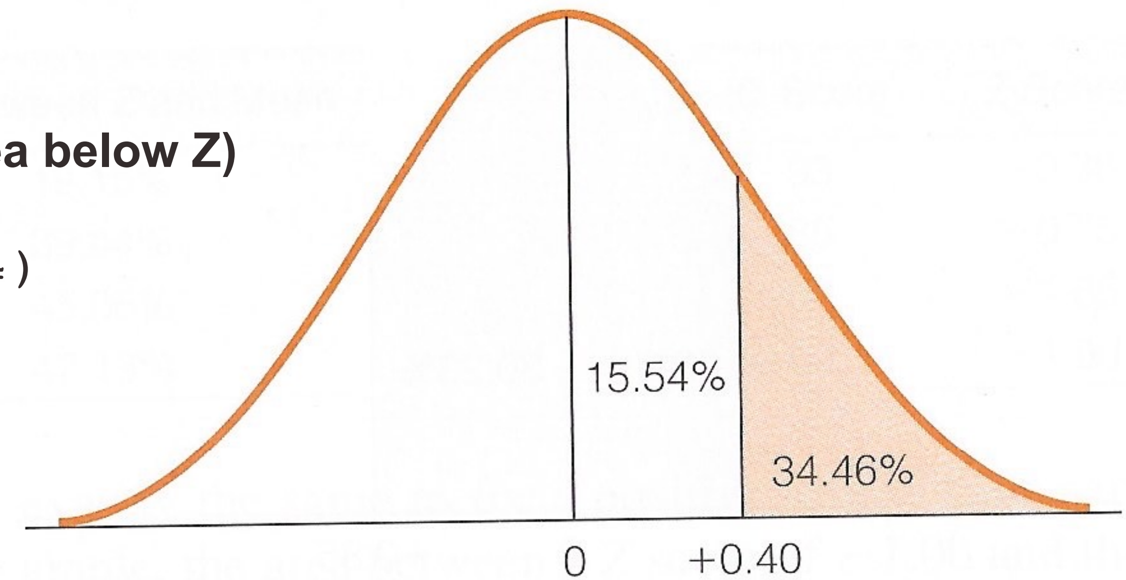
Area above Z = 0.40

- Finding the area above a positive Z score
  - Z = +0.40
  - Area from column c = 0.3446 or 34.46%

Command in Stata (normal shows area below Z)

```
di 1-normal(0.4)

.34457826
```

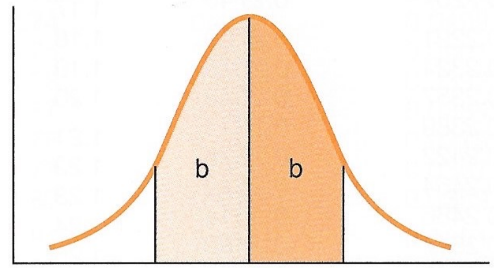


Source: Healey 2015, p.130.

Negative score

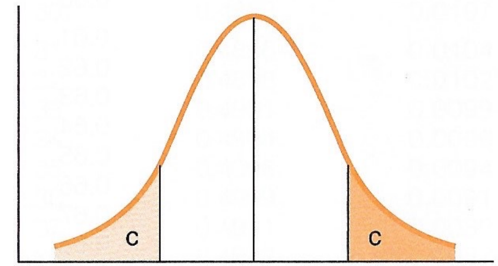
- Find your Z score in Column A
- To find area below a negative score
  - Look in column c
- To find area above a negative score
  - Add column b area to 0.50

FIGURE A.1 Area Between Mean and *Z*



(a) *Z*	(b) Area Between Mean and *Z*	(c) Area Beyond *Z*
0.00	0.0000	0.5000
0.01	0.0040	0.4960
0.02	0.0080	0.4920
0.03	0.0120	0.4880
0.04	0.0160	0.4840
0.05	0.0199	0.4801
0.06	0.0239	0.4761
0.07	0.0279	0.4721
0.08	0.0319	0.4681
0.09	0.0359	0.4641
0.10	0.0398	0.4602
0.11	0.0438	0.4562
0.12	0.0478	0.4522
0.13	0.0517	0.4483
0.14	0.0557	0.4443
0.15	0.0596	0.4404
0.16	0.0636	0.4364
0.17	0.0675	0.4325
0.18	0.0714	0.4286
0.19	0.0753	0.4247
0.20	0.0793	0.4207

FIGURE A.2 Area Beyond *Z*



(a) *Z*	(b) Area Between Mean and *Z*	(c) Area Beyond *Z*
0.21	0.0832	0.4168
0.22	0.0871	0.4129
0.23	0.0910	0.4090
0.24	0.0948	0.4052
0.25	0.0987	0.4013
0.26	0.1026	0.3974
0.27	0.1064	0.3936
0.28	0.1103	0.3897
0.29	0.1141	0.3859
0.30	0.1179	0.3821
0.31	0.1217	0.3783
0.32	0.1255	0.3745
0.33	0.1293	0.3707
0.34	0.1331	0.3669
0.35	0.1368	0.3632
0.36	0.1406	0.3594
0.37	0.1443	0.3557
0.38	0.1480	0.3520
0.39	0.1517	0.3483
0.40	0.1554	0.3446
...	...	...

Source: Healey 2015, Appendix A, p.443.

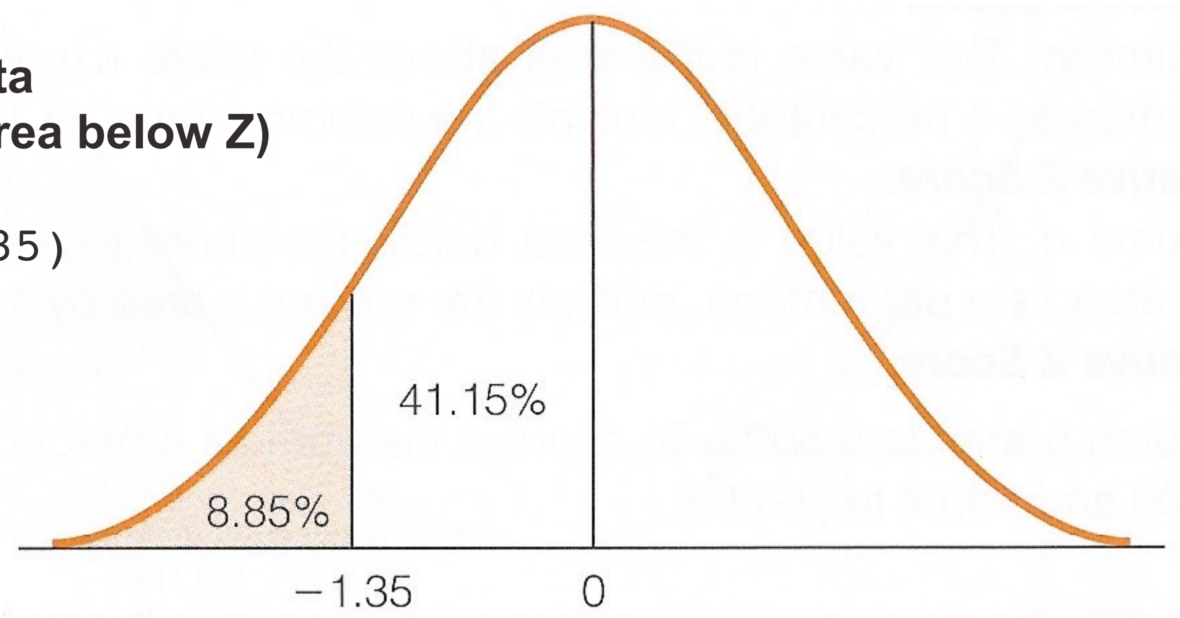
Area below Z = –1.35

- Finding the area below a negative Z score
  - Z = –1.35
  - Area from column c = 0.0885 or 8.85%

Command in Stata
(normal shows area below Z)

```
di normal(-1.35)

.08850799
```



Source: Healey 2015, p.129.

Between scores, opposite sides of mean

- Find your Z scores in Column A
- To find area between two scores on opposite sides of the mean
  - Find the areas between each score and the mean from column b
  - Add the two areas

FIGURE A.1 Area Between Mean and *Z*

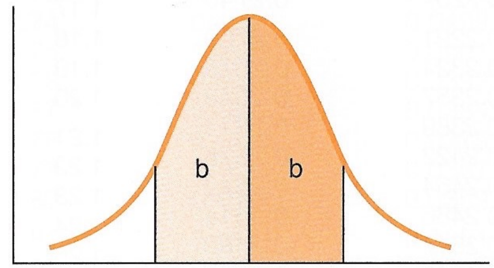
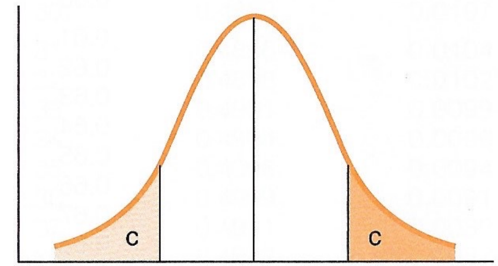


FIGURE A.2 Area Beyond *Z*



(a) *Z*	(b) Area Between Mean and *Z*	(c) Area Beyond *Z*
0.00	0.0000	0.5000
0.01	0.0040	0.4960
0.02	0.0080	0.4920
0.03	0.0120	0.4880
0.04	0.0160	0.4840
0.05	0.0199	0.4801
0.06	0.0239	0.4761
0.07	0.0279	0.4721
0.08	0.0319	0.4681
0.09	0.0359	0.4641
0.10	0.0398	0.4602
0.11	0.0438	0.4562
0.12	0.0478	0.4522
0.13	0.0517	0.4483
0.14	0.0557	0.4443
0.15	0.0596	0.4404
0.16	0.0636	0.4364
0.17	0.0675	0.4325
0.18	0.0714	0.4286
0.19	0.0753	0.4247
0.20	0.0793	0.4207

(a) *Z*	(b) Area Between Mean and *Z*	(c) Area Beyond *Z*
0.21	0.0832	0.4168
0.22	0.0871	0.4129
0.23	0.0910	0.4090
0.24	0.0948	0.4052
0.25	0.0987	0.4013
0.26	0.1026	0.3974
0.27	0.1064	0.3936
0.28	0.1103	0.3897
0.29	0.1141	0.3859
0.30	0.1179	0.3821
0.31	0.1217	0.3783
0.32	0.1255	0.3745
0.33	0.1293	0.3707
0.34	0.1331	0.3669
0.35	0.1368	0.3632
0.36	0.1406	0.3594
0.37	0.1443	0.3557
0.38	0.1480	0.3520
0.39	0.1517	0.3483
0.40	0.1554	0.3446
...	...	...

Source: Healey 2015, Appendix A, p.443.

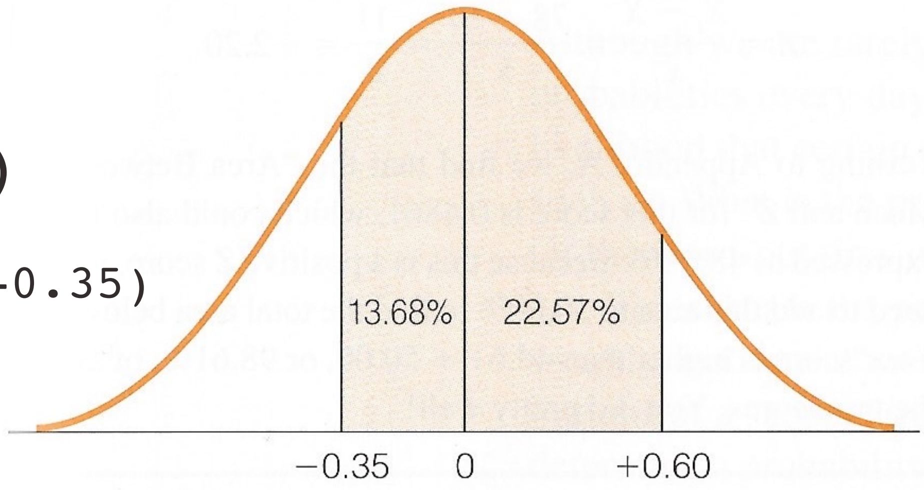
Area between two scores, opposite sides of mean

- Finding the area between Z scores on different sides of the mean
  - Z = –0.35, area from column b = 0.1368
  - Z = +0.60, area from column b = 0.2257
  - Area = 0.1368 + 0.2257 = 0.3625 or 36.25%

Command in Stata
(normal shows area below Z)

```
di normal(0.6)-normal(-0.35)

.36257753
```



Source: Healey 2015, p.131.

Between scores, same side of mean

- Find your Z scores in Column A
- To find area between two scores on the same side of the mean
  - Find the area between each score and the mean from column b
  - Subtract the smaller area from the larger area

FIGURE A.1 Area Between Mean and *Z*

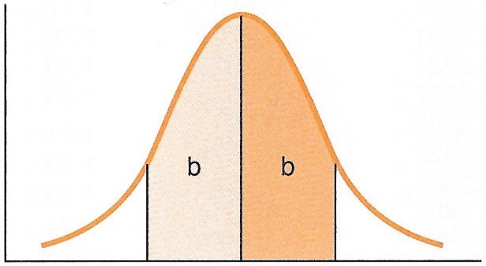
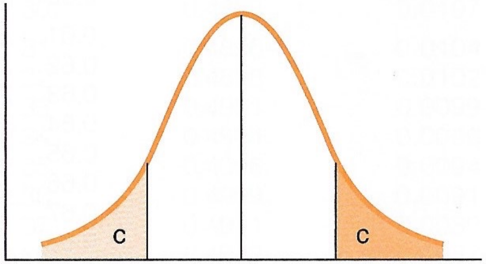


FIGURE A.2 Area Beyond *Z*



(a) *Z*	(b) Area Between Mean and *Z*	(c) Area Beyond *Z*
0.00	0.0000	0.5000
0.01	0.0040	0.4960
0.02	0.0080	0.4920
0.03	0.0120	0.4880
0.04	0.0160	0.4840
0.05	0.0199	0.4801
0.06	0.0239	0.4761
0.07	0.0279	0.4721
0.08	0.0319	0.4681
0.09	0.0359	0.4641
0.10	0.0398	0.4602
0.11	0.0438	0.4562
0.12	0.0478	0.4522
0.13	0.0517	0.4483
0.14	0.0557	0.4443
0.15	0.0596	0.4404
0.16	0.0636	0.4364
0.17	0.0675	0.4325
0.18	0.0714	0.4286
0.19	0.0753	0.4247
0.20	0.0793	0.4207

(a) *Z*	(b) Area Between Mean and *Z*	(c) Area Beyond *Z*
0.21	0.0832	0.4168
0.22	0.0871	0.4129
0.23	0.0910	0.4090
0.24	0.0948	0.4052
0.25	0.0987	0.4013
0.26	0.1026	0.3974
0.27	0.1064	0.3936
0.28	0.1103	0.3897
0.29	0.1141	0.3859
0.30	0.1179	0.3821
0.31	0.1217	0.3783
0.32	0.1255	0.3745
0.33	0.1293	0.3707
0.34	0.1331	0.3669
0.35	0.1368	0.3632
0.36	0.1406	0.3594
0.37	0.1443	0.3557
0.38	0.1480	0.3520
0.39	0.1517	0.3483
0.40	0.1554	0.3446
...	...	...

Source: Healey 2015, Appendix A, p.443.

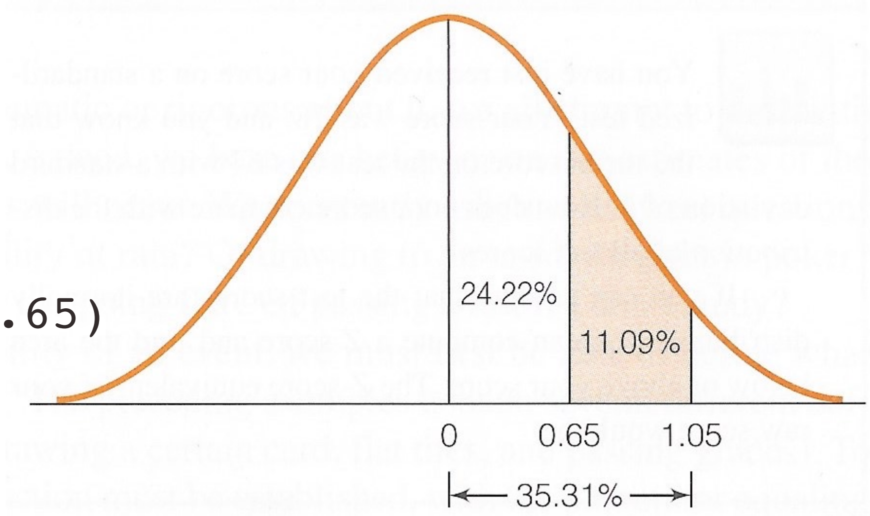
Area between two scores, same side of mean

- Finding the area between Z scores on the same side of the mean
  - Z = +0.65, area from column b = 0.2422
  - Z = +1.05, area from column b = 0.3531
  - Area = 0.3531 – 0.2422 = 0.1109 or 11.09%

Command in Stata (normal shows area below Z)

```
di normal(1.05)-normal(0.65)

.11098705
```



Source: Healey 2015, p.131.

Estimating probabilities

- Areas under the curve can also be expressed as probabilities
- Probabilities are proportions
  - They range from 0.00 to 1.00
- The higher the value, the greater the probability
  - The more likely the event

Example

- If a distribution has mean equals to 13 and standard deviation equals to 4

- What is the probability of randomly selecting a score of 19 or more?

$$Z = \frac{X_i - \bar{X}}{s} = \frac{19 - 13}{4} = \frac{6}{4} = 1.5$$

- Command in Stata (normal shows area below Z)

```
di 1-normal(1.5)
```

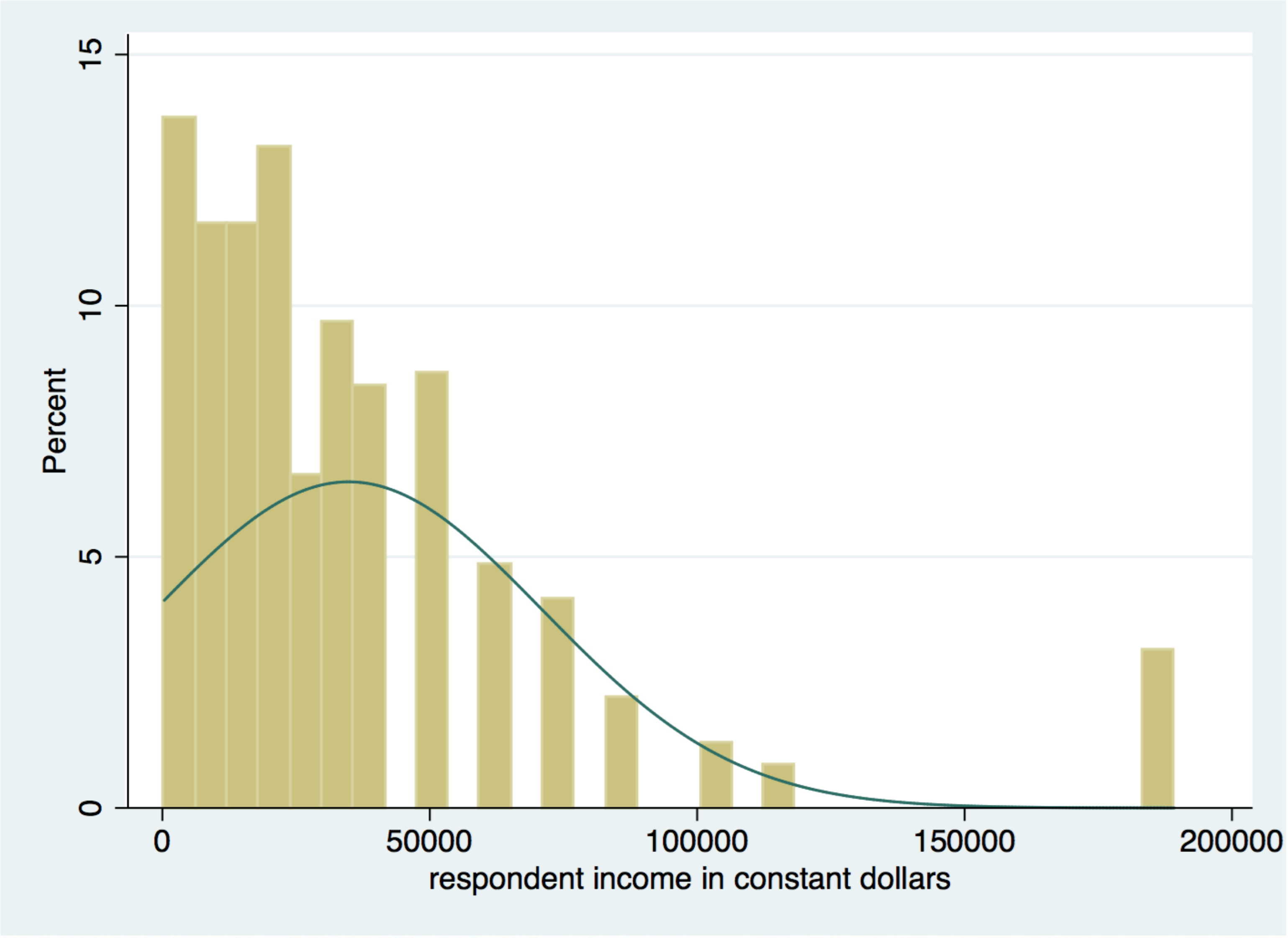
p = 0.0668072

Determining normality

- Some statistical methods require random selection of respondents from a population with normal distribution for its variables

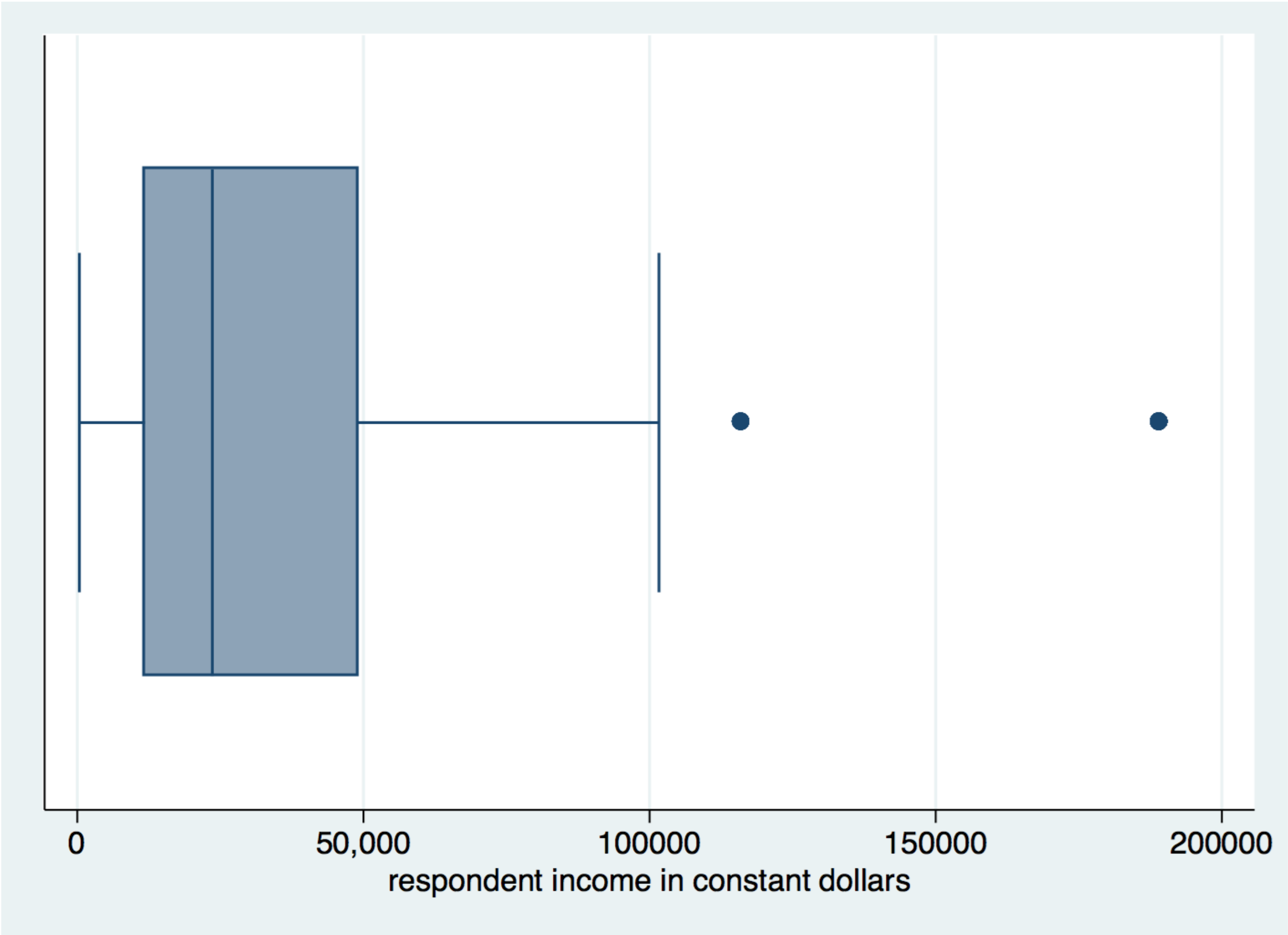
- We can analyze histograms, boxplots, outliers, quantile-normal plots to determine if variables have a normal distribution

Histogram of income



Source: 2016 General Social Survey.

Boxplot of income

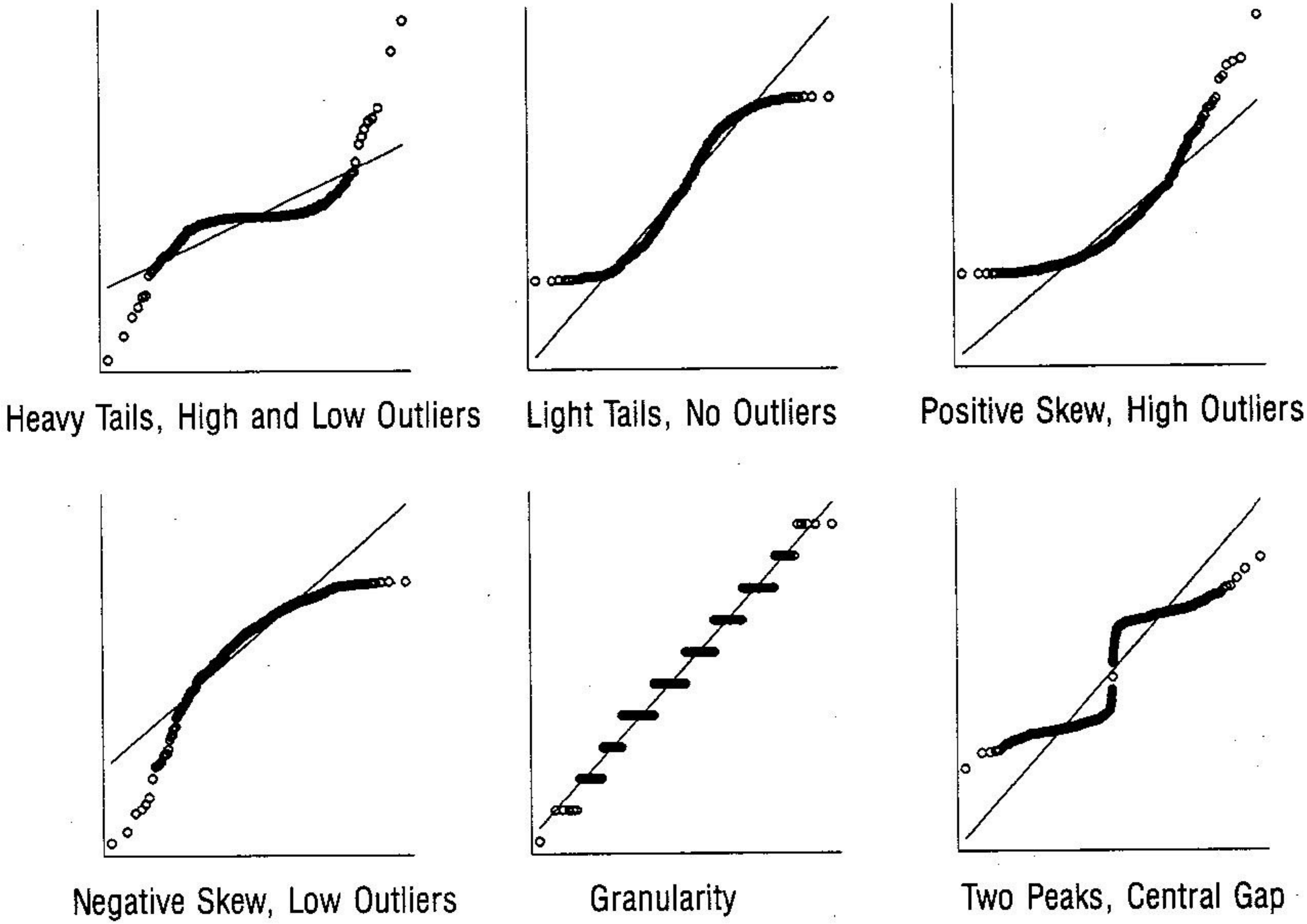


Source: 2016 General Social Survey.

Quantile-normal plots

- A quantile-normal plot is a scatter plot
  - One axis has quantiles of the original data
  - The other axis has quantiles of the normal distribution
- If the points do not form a straight line or if the points have a non-linear symmetric pattern
  - The variable does not have a normal distribution
- If the pattern of points is roughly straight
  - The variable has a distribution close to normal
- If the variable has a normal distribution
  - The points would exactly overlap the diagonal line

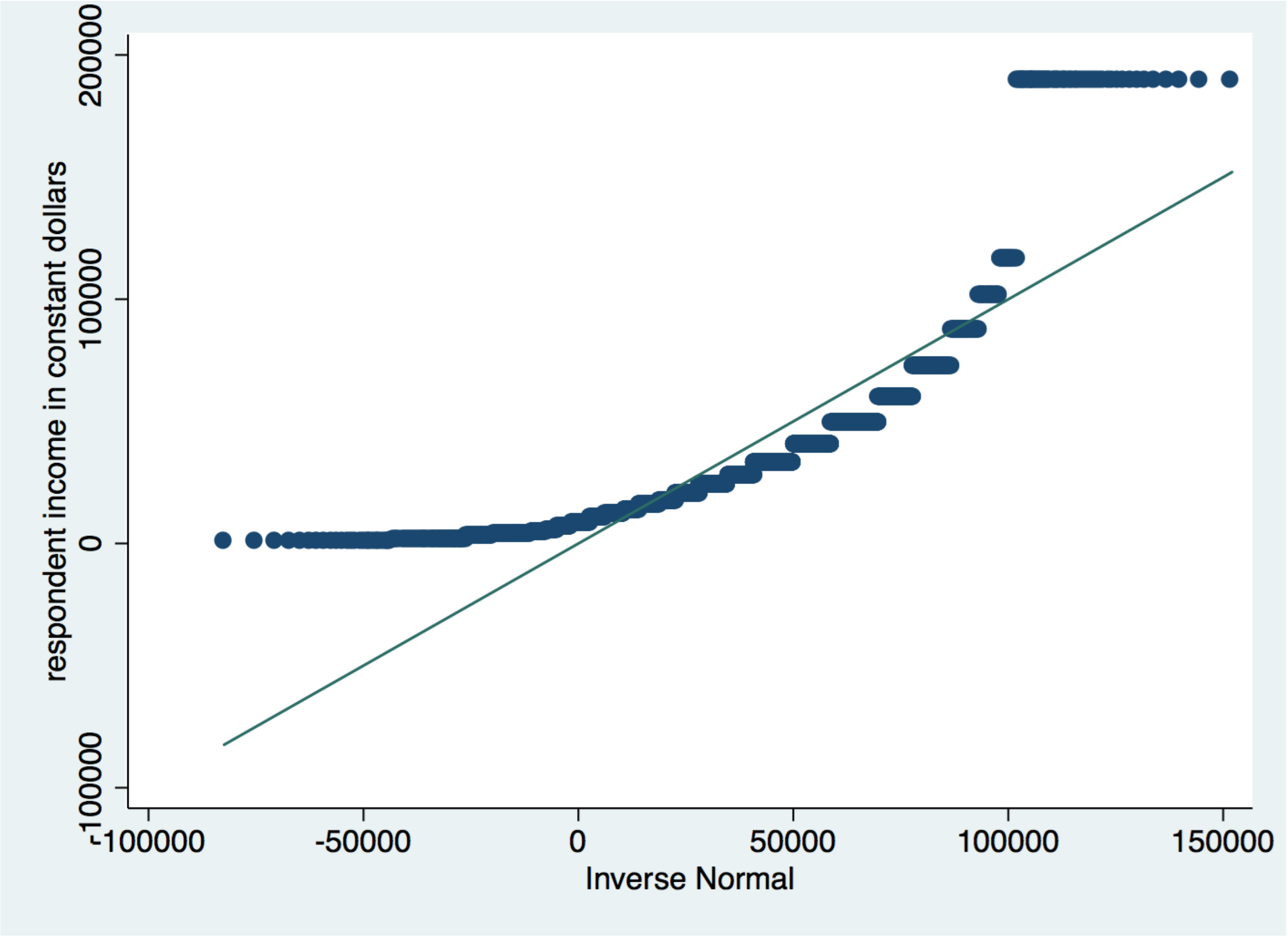
Quantile-normal plots reflect distribution shapes



(discrete values) **(bimodal)**

Source: Hamilton 1992, p.16.

Quantile-normal plot of income



Source: 2016 General Social Survey.

Power transformation

- Lawrence Hamilton ("Regression with Graphics", 1992, p.18–19)

$$Y^3 \longrightarrow q = 3$$

$$Y^2 \longrightarrow q = 2$$

$$Y^1 \longrightarrow q = 1$$

$$Y^{0.5} \longrightarrow q = 0.5$$

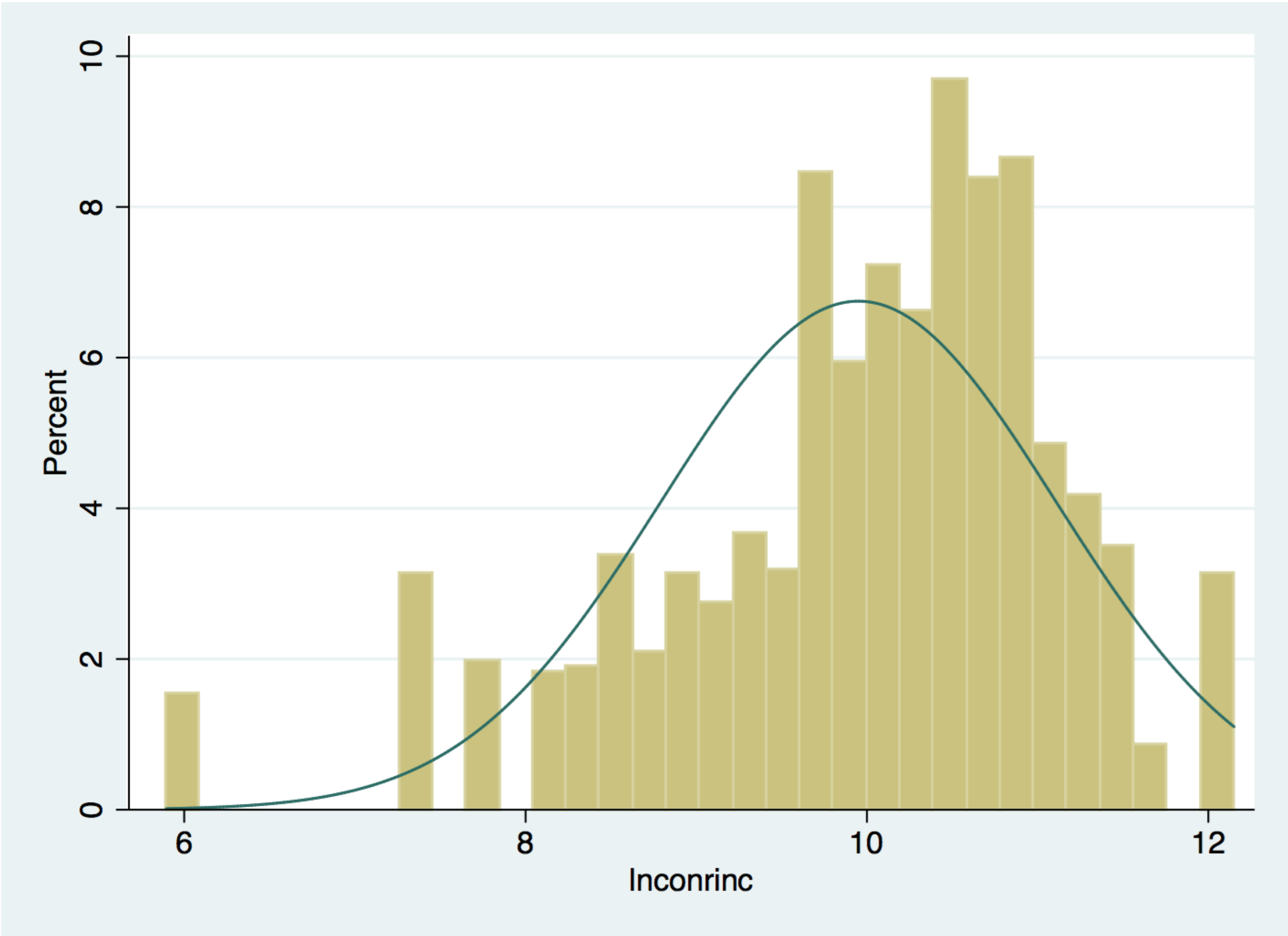
$$\log(Y) \longrightarrow q = 0$$

$$-(Y^{-0.5}) \longrightarrow q = -0.5$$

$$-(Y^{-1}) \longrightarrow q = -1$$

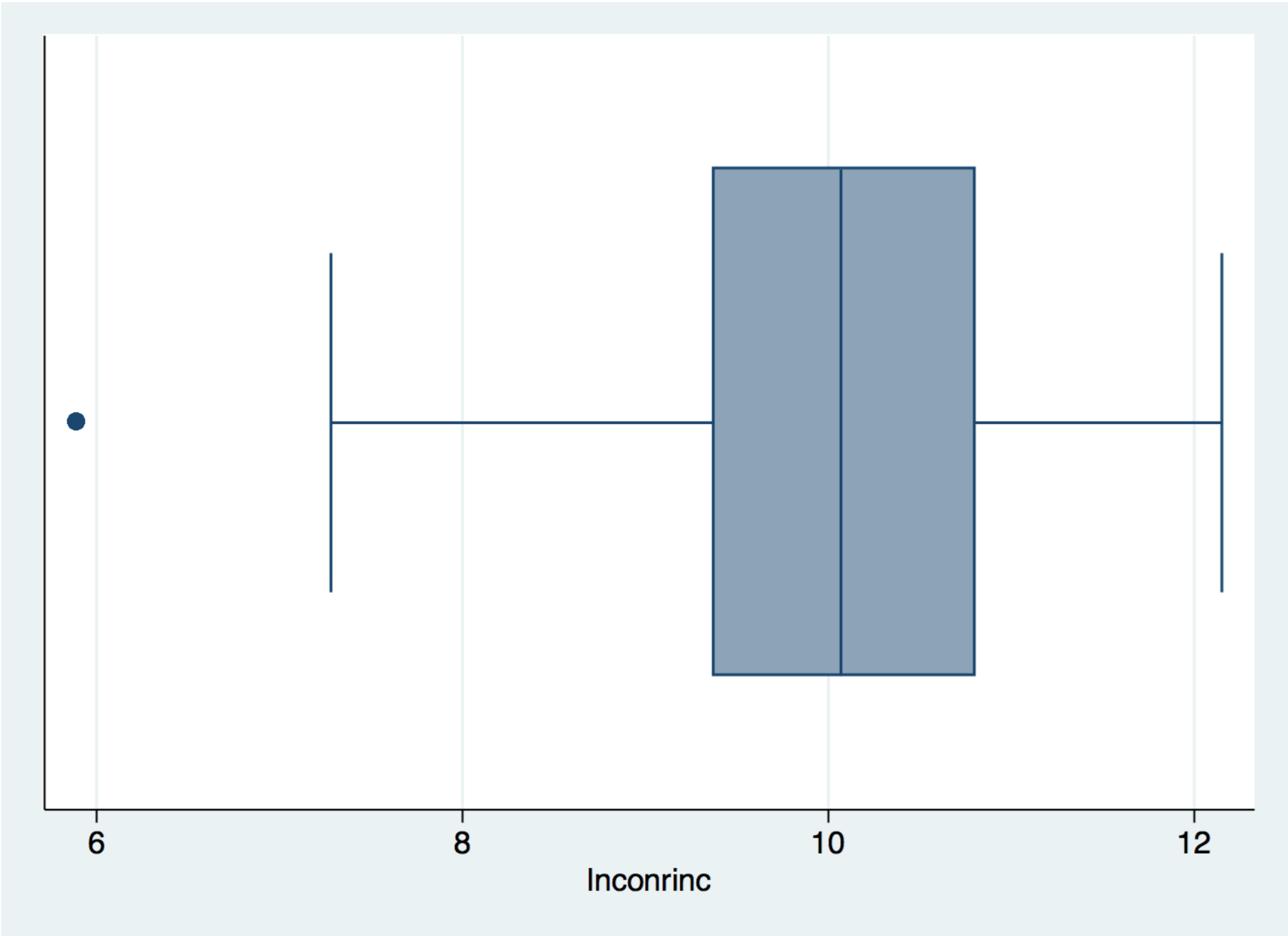
- $q>1$: reduce concentration on the right (reduce negative skew)
- $q=1$: original data
- $q<1$: reduce concentration on the left (reduce positive skew)
- $log(x+1)$ may be applied when $x=0$. If distribution of $log(x+1)$ is normal, it is called lognormal distribution

Histogram of log of income



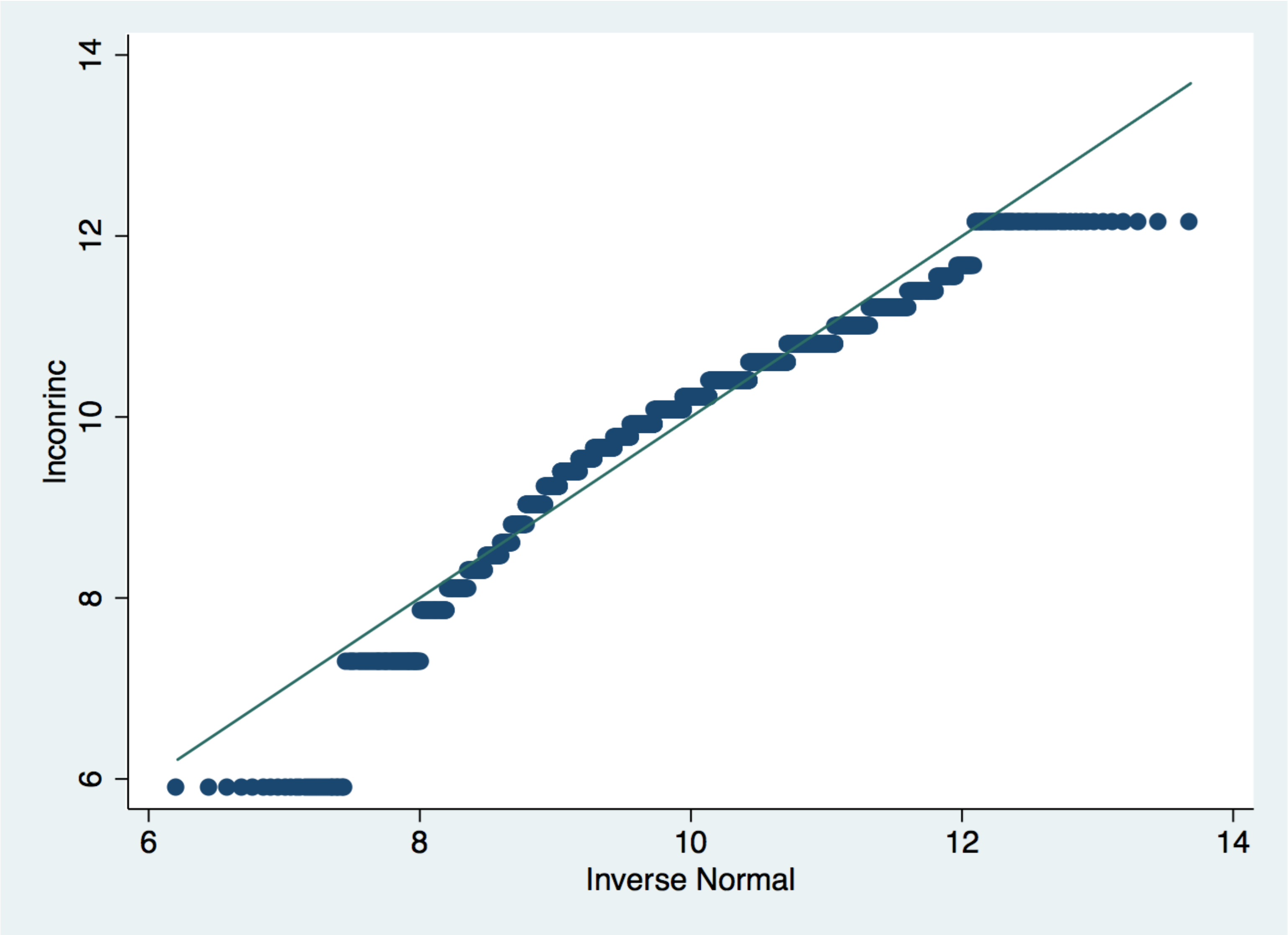
Source: 2016 General Social Survey.

Boxplot of log of income



Source: 2016 General Social Survey.

Quantile-normal plot of log of income

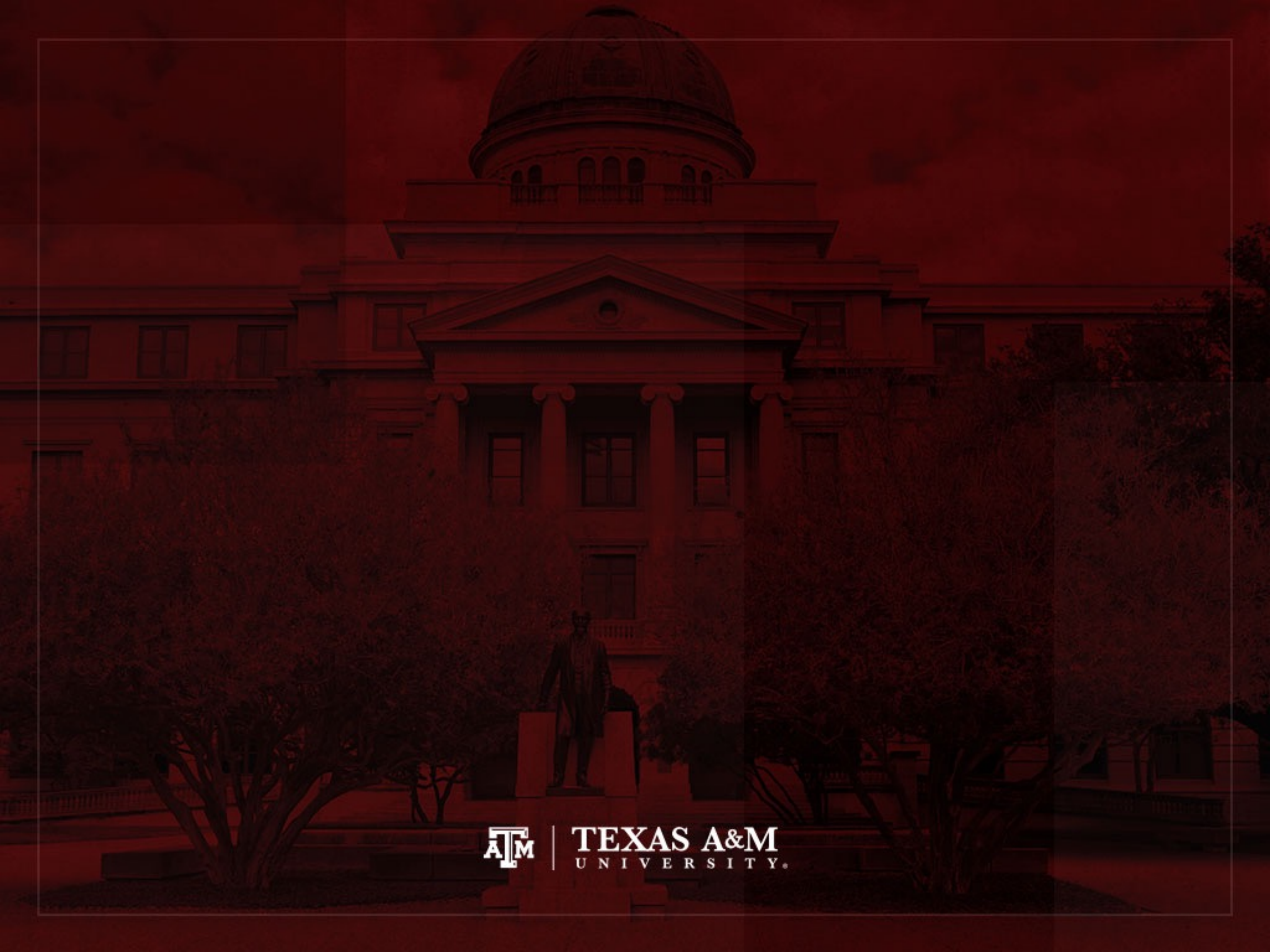


Source: 2016 General Social Survey.

Points to remember

- Cases with scores close to the mean are common and those with scores far from the mean are rare

- The normal curve is essential for understanding inferential statistics in Part II of the textbook


TEXAS A&M
UNIVERSITY

Inferential statistics

- Explain the purpose of inferential statistics in terms of generalizing from a sample to a population
- Define and explain the basic techniques of random sampling
- Explain and define these key terms: population, sample, parameter, statistic, representative, EPSEM sampling techniques
- Differentiate between the sampling distribution, the sample, and the population
- Explain two theorems

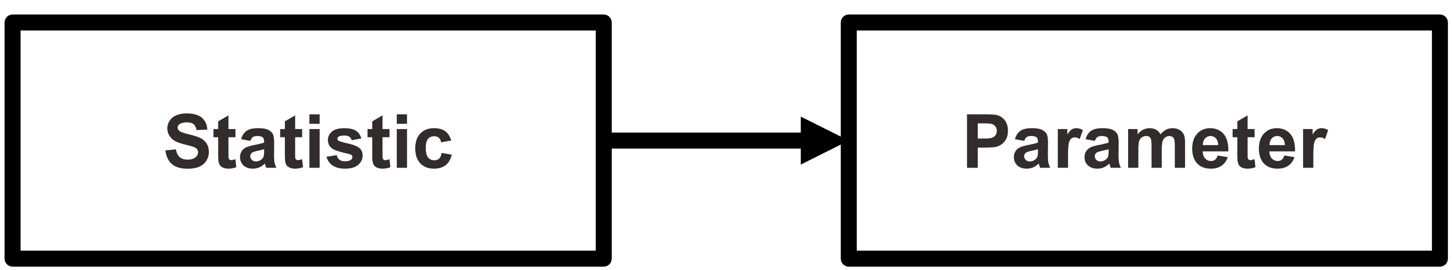
Basic logic and terminology

- **Problem**
- The populations we wish to study are almost always so large that we are unable to gather information from every case

- **Solution**
- We choose a sample – a carefully chosen subset of the population – and use information gathered from the cases in the sample to generalize to the population

Basic logic and terminology

- **Statistics** are mathematical characteristics of samples
- **Parameters** are mathematical characteristics of populations
- **Statistics** are used to estimate **parameters**

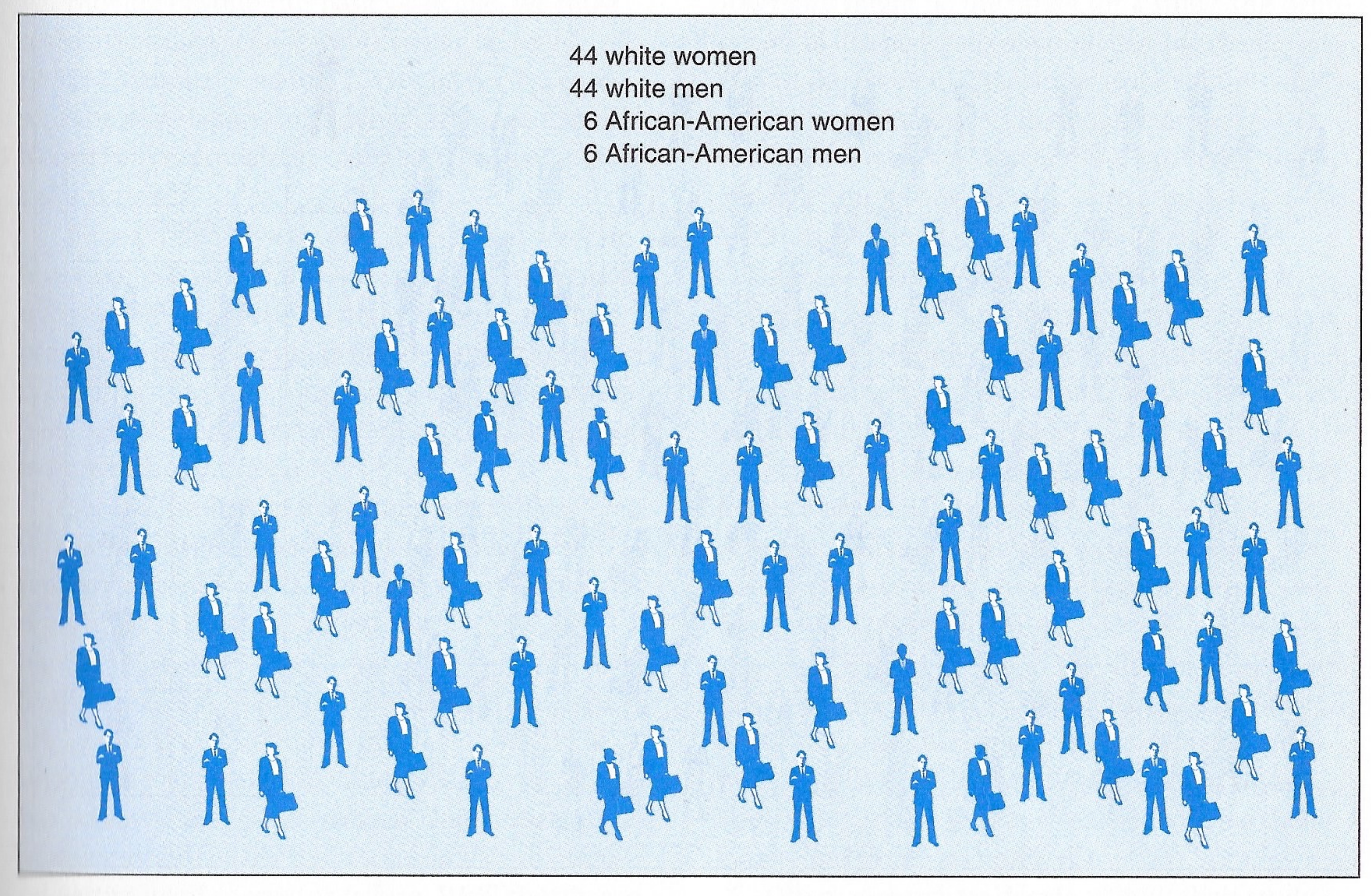


Samples

- Must be representative of the population
  - Representative: The sample has the same characteristics as the population

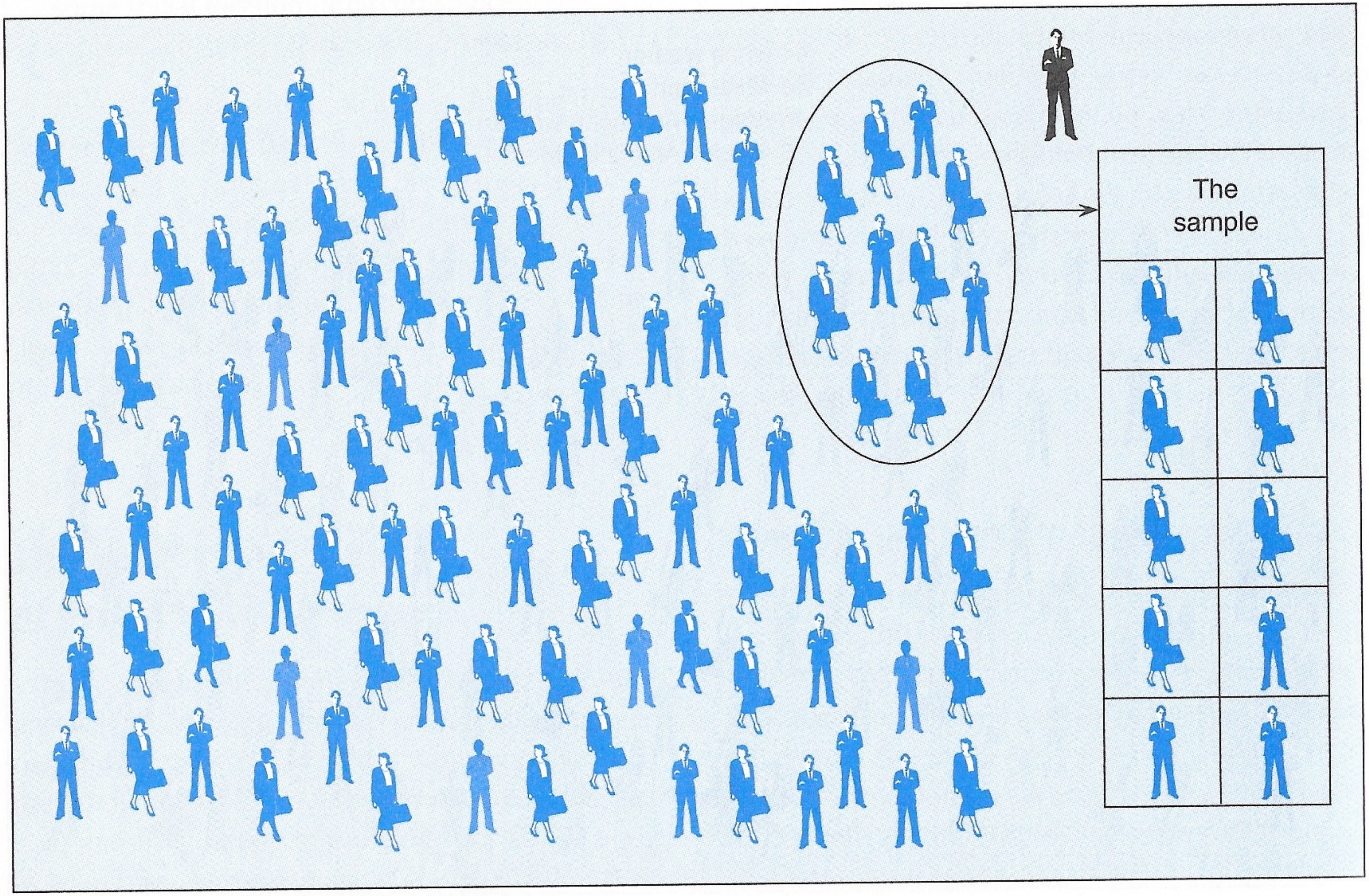
- How can we ensure samples are representative?
  - Samples drawn according to the rule of **EPSEM** (**e**qual **p**robability of **s**election **m**ethod)
  - If every case in the population has the same chance of being selected, the sample is likely to be representative

A population of 100 people



Source: Babbie 2001, p.183.

Nonprobability sampling



Source: Babbie 2001, p.184.

EPSEM sampling techniques

1. Simple random sampling
2. Systematic sampling
3. Stratified sampling
4. Cluster sampling

1. Simple random sampling

- To begin, we need
  - A list of the population
- Then, we need a method for selecting cases from the population, so each case has the same probability of being selected
  - The principle of EPSEM
  - A sample selected this way is very likely to be representative of the population
  - Variable in population should have a normal distribution or n>30

Example

- You want to know what percent of students at a large university work during the semester
- Draw a sample size (*n*) of 500 from a list of all students (*N*=20,000)
- Assume the list is available from the Registrar
- How can you draw names, so every student has the same chance of being selected?

Example

- Each student has a unique, 6 digit ID number that ranges from 000001 to 999999
- Use a table of random numbers or a computer program to select 500 ID numbers with 6 digits each
- Each time a randomly selected 6 digit number matches the ID of a student, that student is selected for the sample
- Continue until 500 names are selected

Example

- **Stata**

```
set obs 500
generate student = runiformint(1,999999)
sum student
    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
     student |        500    482562.6    283480.9       3652     997200
```

- **Excel**
  - RANDBETWEEN (minimum,maximum)
    - Returns a random number between those you specify
    - Drag the function to 500 cells

=RANDBETWEEN(1,999999)

  - RANDARRAY (rows,columns,minimum,maximum)

=RANDARRAY(500,1,1,999999)

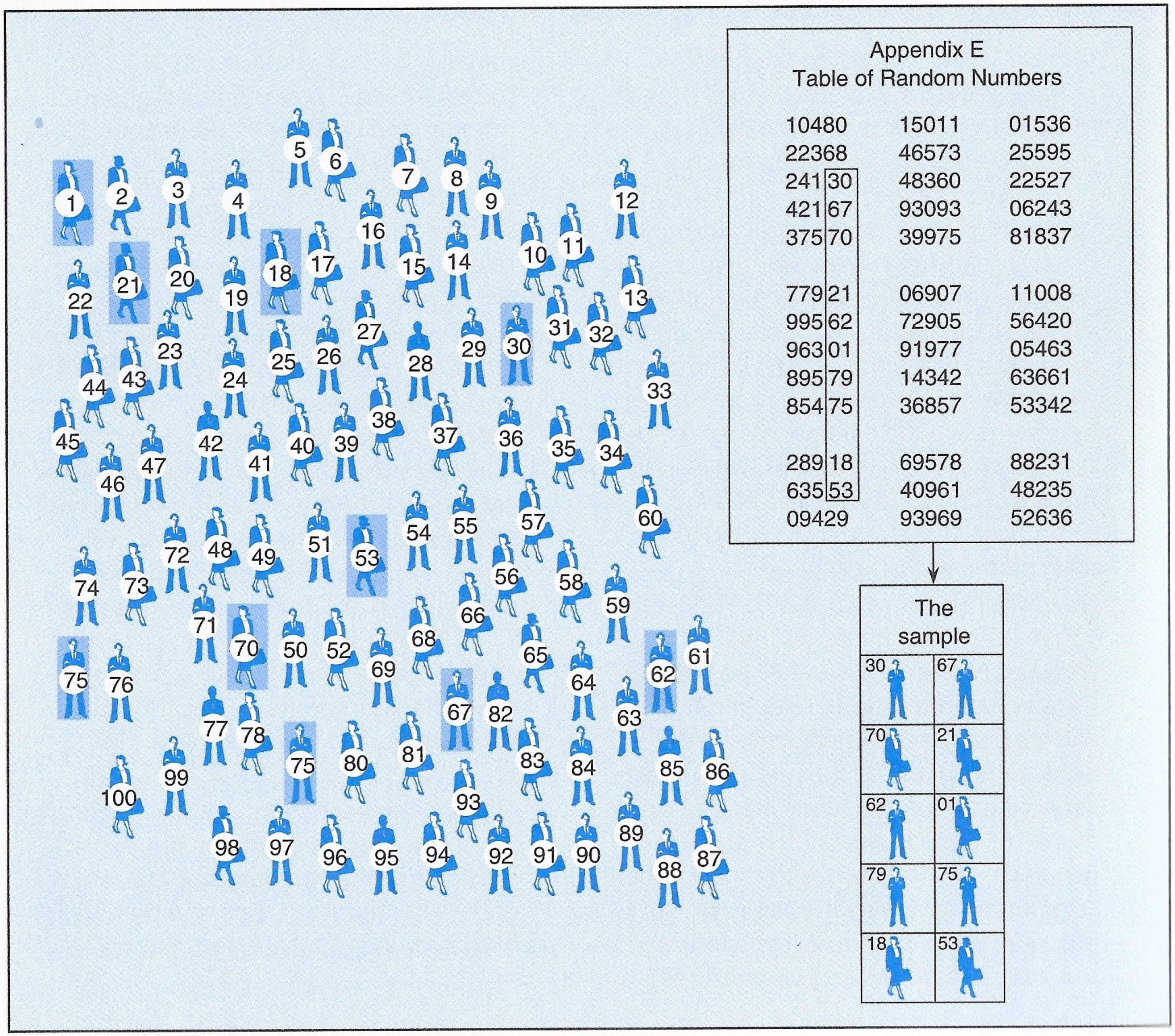
Example

- Disregard duplicate numbers
- Ignore cases in which no student ID matches the randomly selected number
- After questioning each of these 500 students, you find that 368 (74%) work during the semester

Applying logic and terminology

- In the previous example:
- **Population:** All 20,000 students
- **Sample:** 500 students selected and interviewed
- **Statistic:** 74% (percentage of sample that held a job during the semester)
- **Parameter:** Percentage of all students in the population who held a job

Simple random sample



Source: Babbie 2001, p.200.

2. Systematic sampling

- Useful for large populations
- Randomly select the first case then select every k^{th} case
- **Sampling interval**
  - Distance between elements selected in the sample
  - Population size (*N*) divided by sample size (*n*)
- **Sampling ratio**
  - Proportion of selected elements in the population
  - Sample size (*n*) divided by population size (*N*)
- Can be problematic if the list of cases is not truly random or demonstrates some patterning

Source: Babbie 2001, p.197–198.

Example

- If a list contained 10,000 elements and we want a sample of 1,000
- Sampling interval
  - Population size / sample size = 10,000 / 1,000 = 10
  - We would select every 10th element for our sample
- Sampling ratio
  - Sample size / population size = 1,000 / 10,000 = 1/10
  - Proportion of selected elements in population
- Select the first element at random

Source: Babbie 2001, p.197–198.

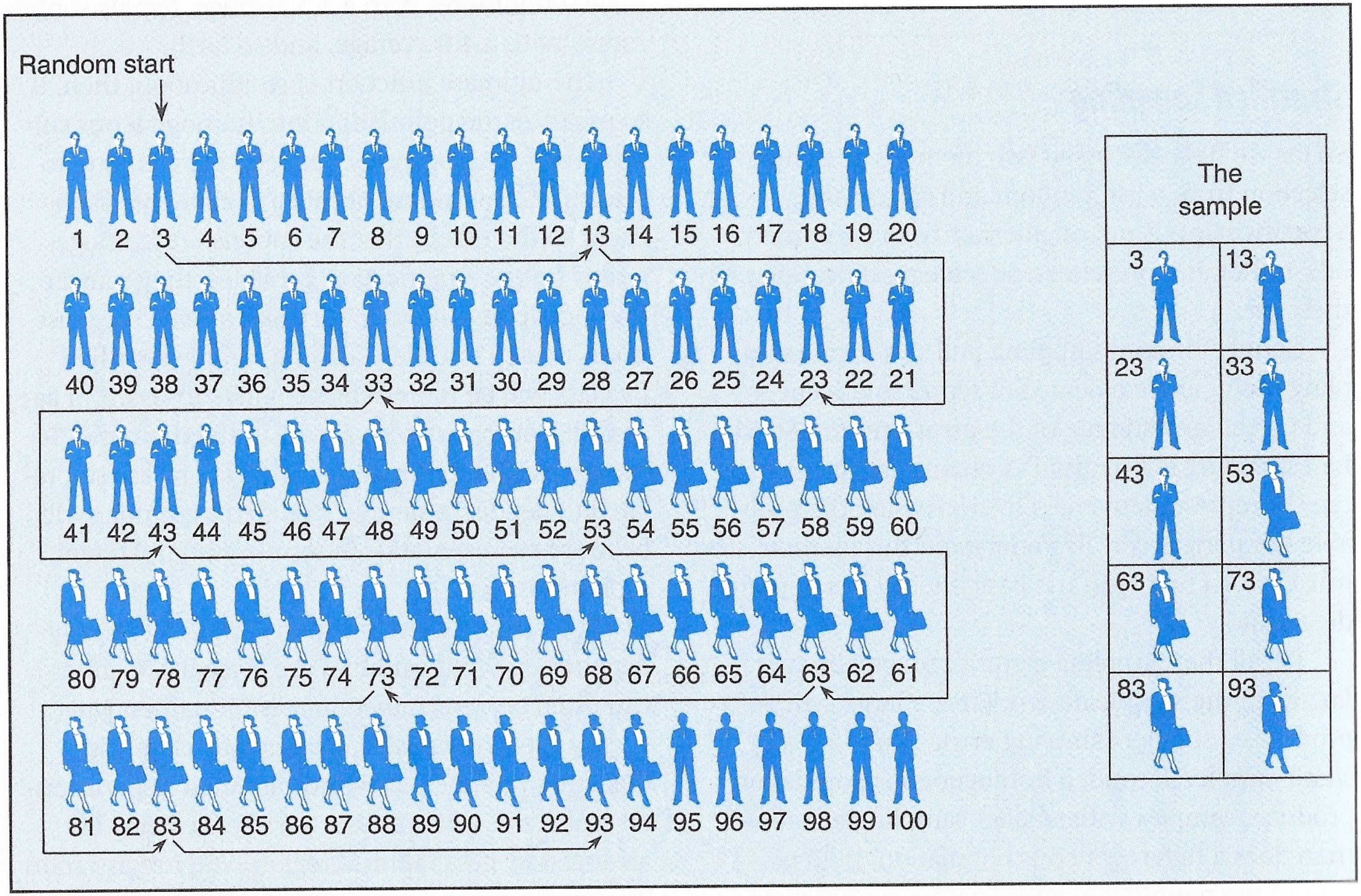
3. Stratified sampling

- It guarantees the sample will be representative on the selected (stratifying) variables
  - Stratification variables relate to research interests
- First, divide the population list into subsets, according to some relevant variable
  - **Homogeneity within subsets**
    - E.g., only women in a subset; only men in another subset
  - **Heterogeneity between subsets**
    - E.g., subset of women is different than subset of men
- Second, sample from the subsets
  - Select the number of cases from each subset proportional to the population

Example

- If you want a sample of 1,000 students
  - That would be representative to the population of students by sex and GPA
- You need to know the population composition
  - E.g., women with a 4.0 average compose 15 percent of the student population
- Your sample should follow that composition
  - In a sample of 1,000 students, you would select 150 women with a 4.0 average

Stratified, systematic sample



Source: Babbie 2001, p.202.

4. Cluster sampling

- Select groups (or clusters) of cases rather than single cases
  - **Heterogeneity within subsets**
    - E.g., each subset has both women and men, following same proportional distribution as population
  - **Homogeneity between subsets**
    - E.g., all subsets with both women and men should be similar
- Clusters are often geographically based
  - For example, cities or voting districts
- Sampling often proceeds in stages
  - Multi-stage cluster sampling
  - Less representative than simple random sampling

Stratified vs. cluster sampling

- **Stratified**
  - Homogeneity within subsets
  - Heterogeneity between subsets
  - Select cases from each subset

Subset of women	Subset of men

- **Cluster**
  - Heterogeneity within subsets (groups, clusters, areas)
  - Homogeneity between subsets
  - Select groups (e.g., area 1) rather than single cases

Area 1: women & men	Area 2: women & men

Sampling distribution

- Sampling distribution is the probabilistic distribution of a statistic for all possible samples of a given size (*n*)
  - It is the distribution of a statistic (e.g., proportion, mean) for all possible outcomes of a certain size
- Central tendency and dispersion
  - Mean is the same as the population mean
  - Standard deviation is referred as standard error
    - It is the population standard deviation divided by the square root of *n*
    - We have to take into account the complex survey design to estimate the standard error (**`svyset`** command in Stata)

Linking sample and population

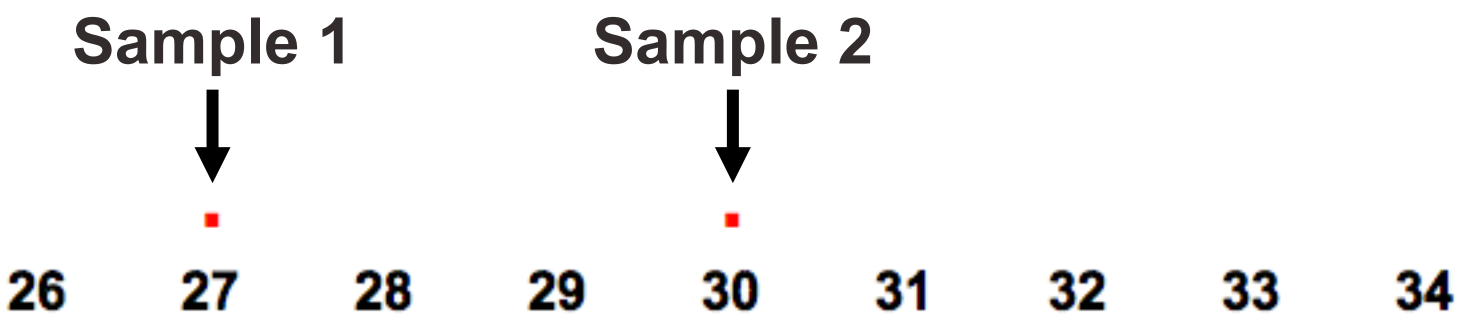
- Every application of inferential statistics involves three different distributions
  - Population: empirical; unknown
  - Sampling distribution: theoretical; known
  - Sample: empirical; known

- In inferential statistics, the sample distribution links the sample with the population

Example

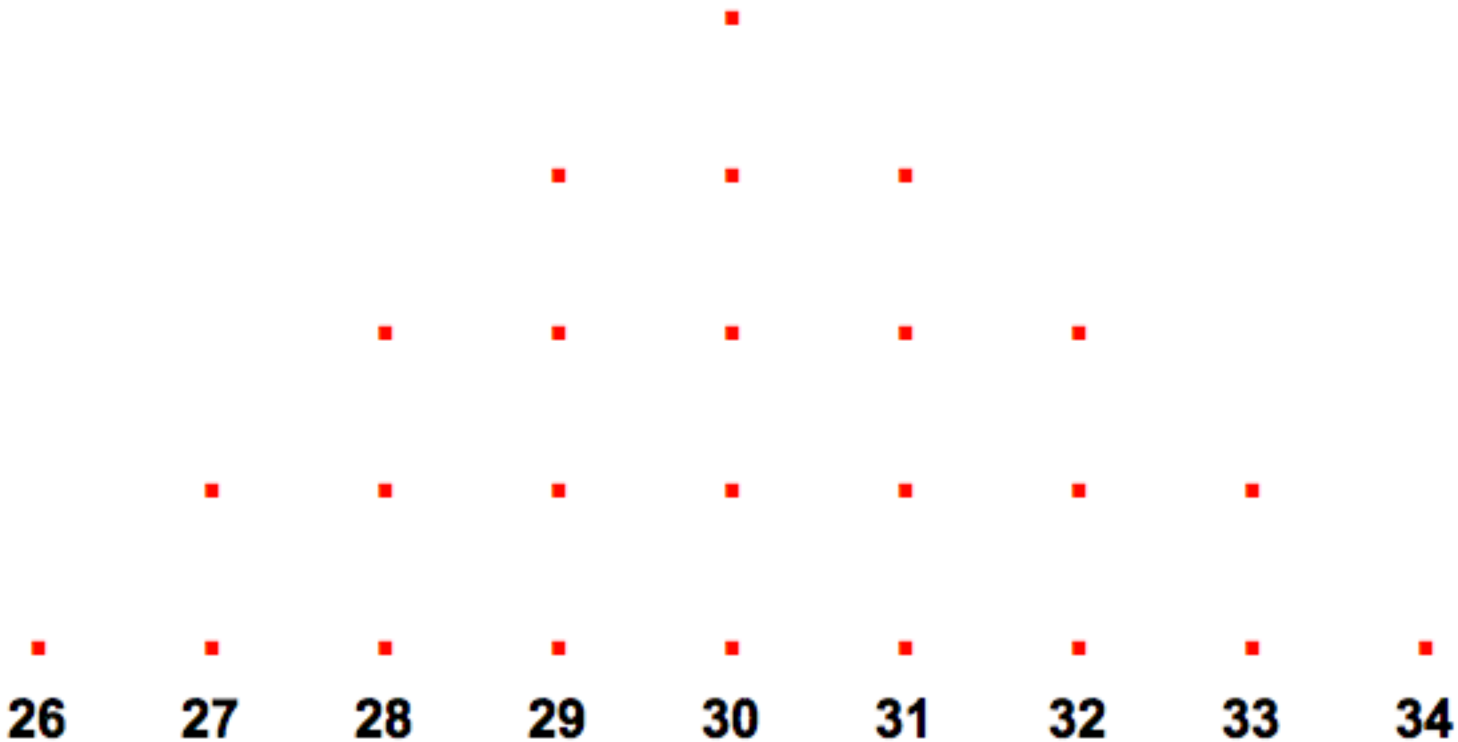
- Suppose we want to gather information on the age of a community of 10,000 individuals
  - Sample 1: *n*=100 people, plot sample's mean of 27
  - Replace people in the sample back to the population

  - Sample 2: *n*=100 people, plot sample's mean of 30
  - Replace people in the sample back to the population

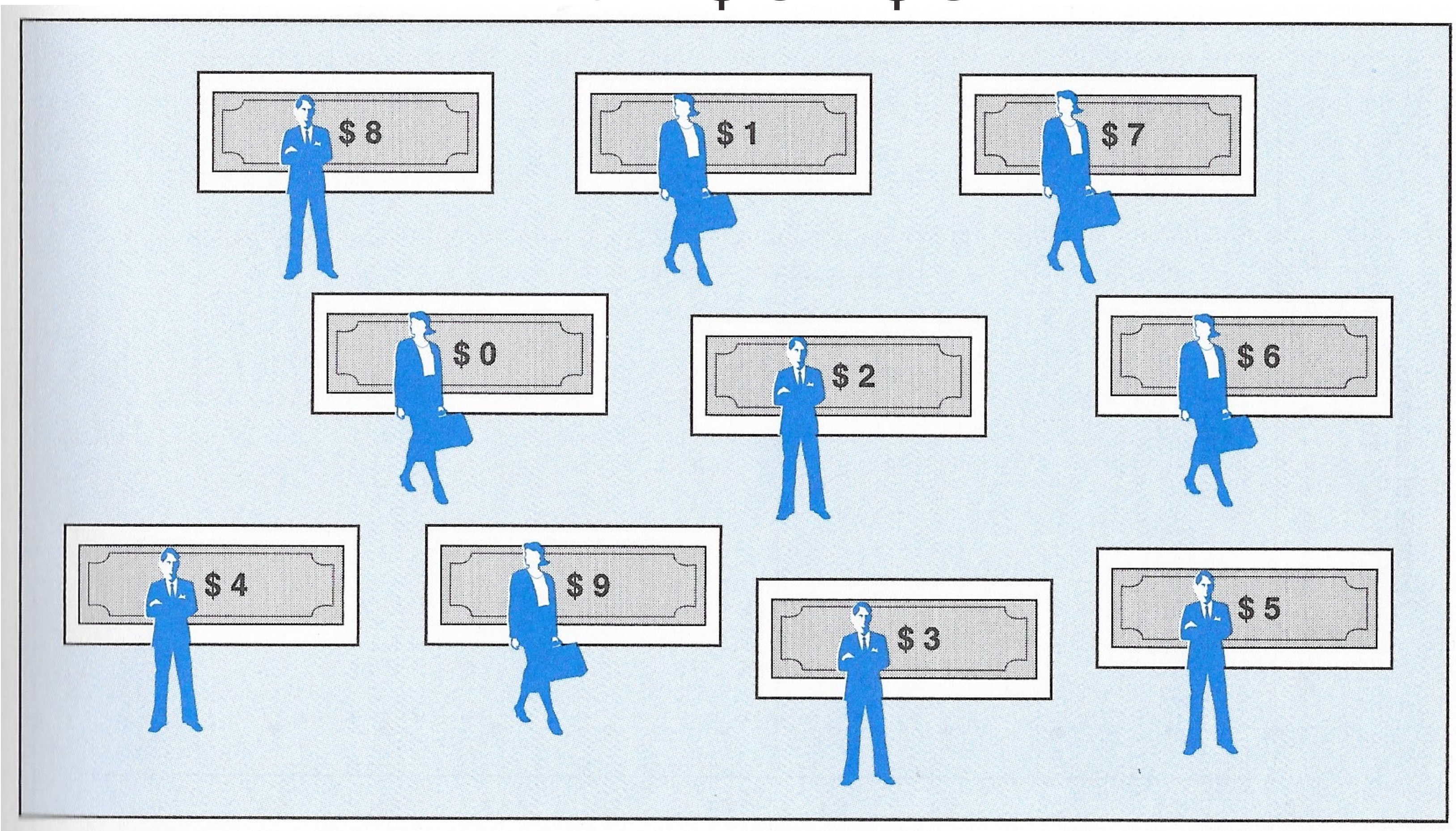


Example

- We repeat this procedure: sampling, replacing
  - Until we have exhausted every possible combination of 100 people from the population of 10,000
  - Sampling distribution has a normal shape

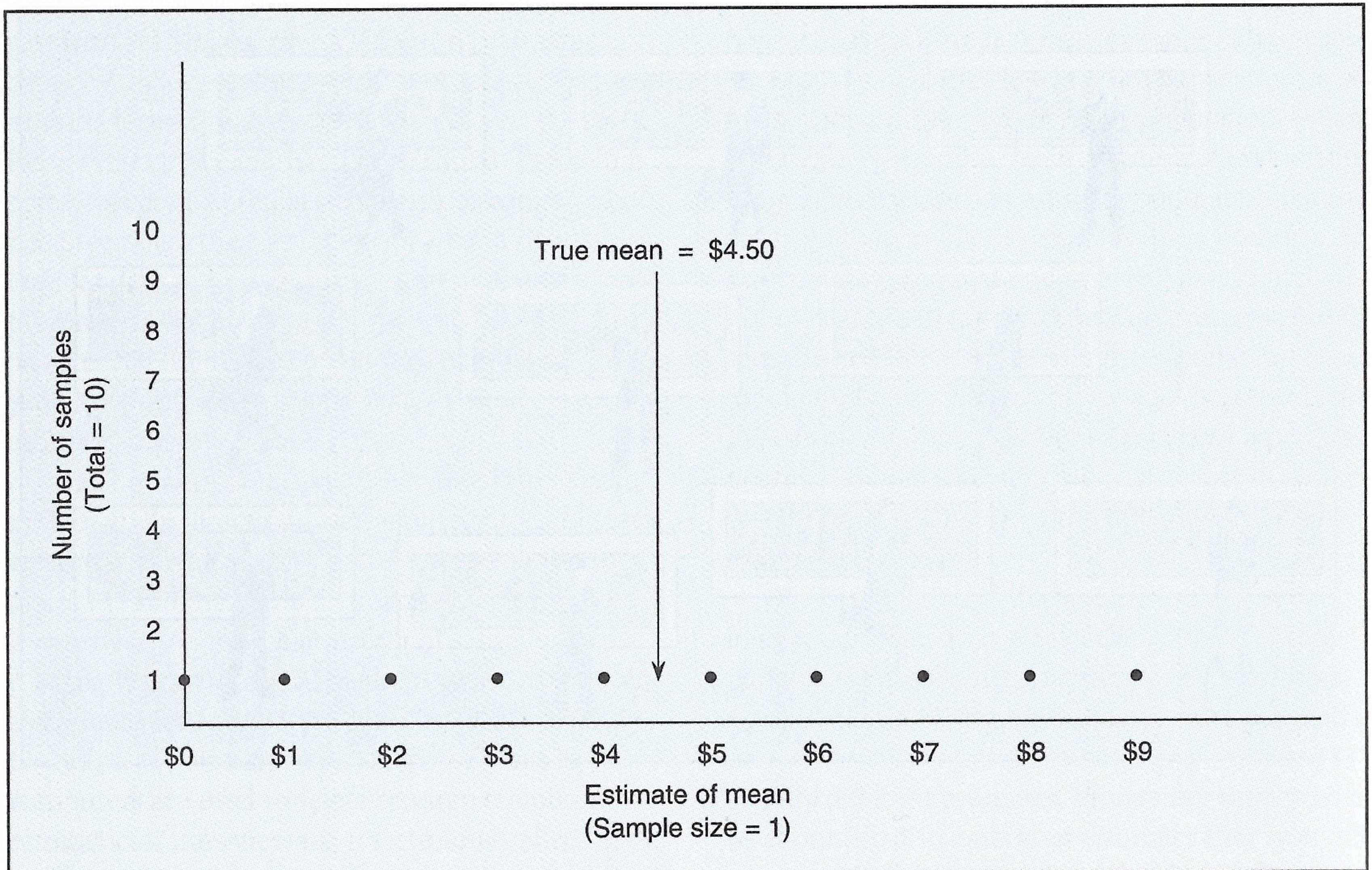


Another example: A population of 10 people with $0–$9



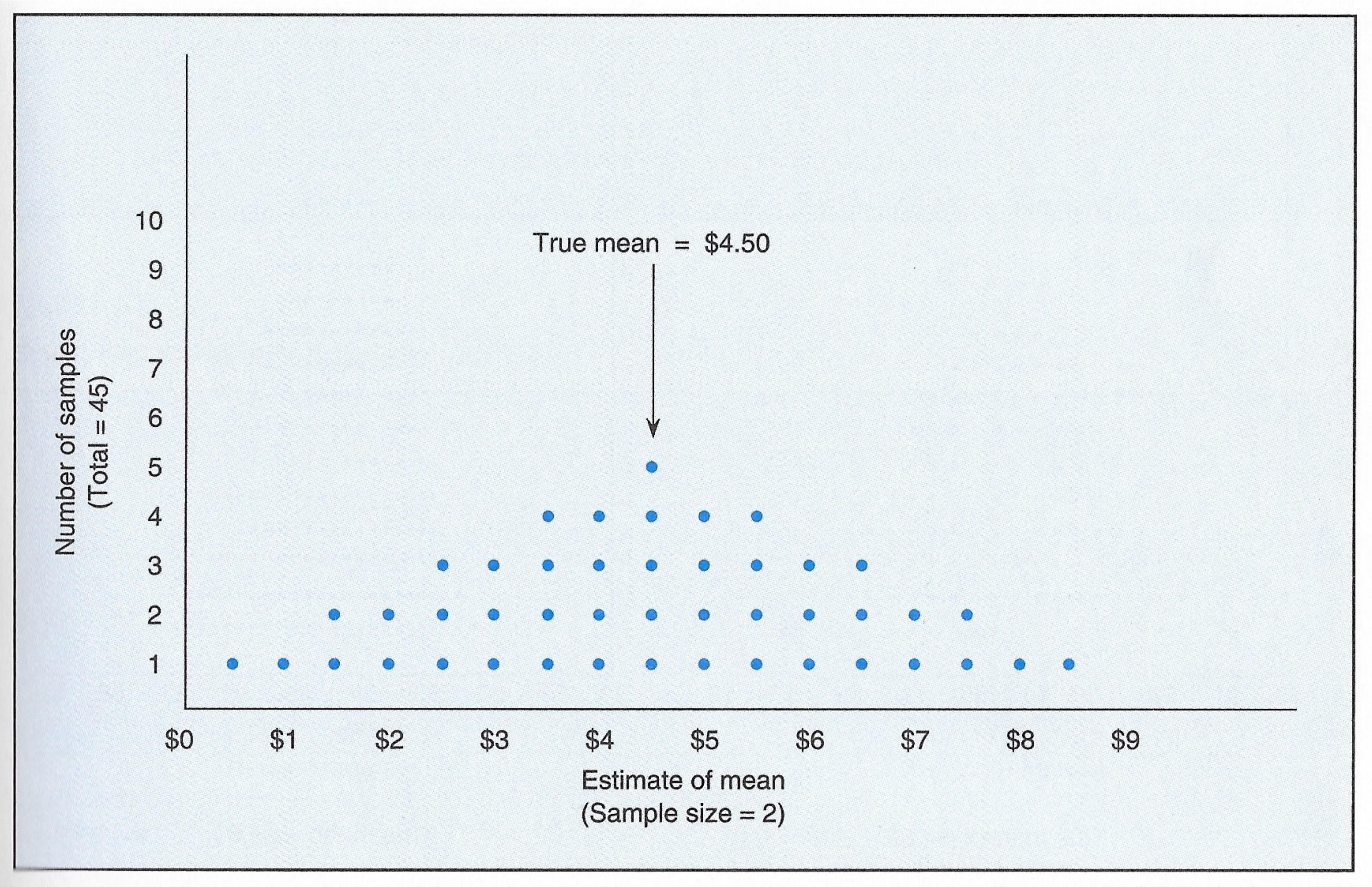
Source: Babbie 2001, p.187.

The sampling distribution (n=1)



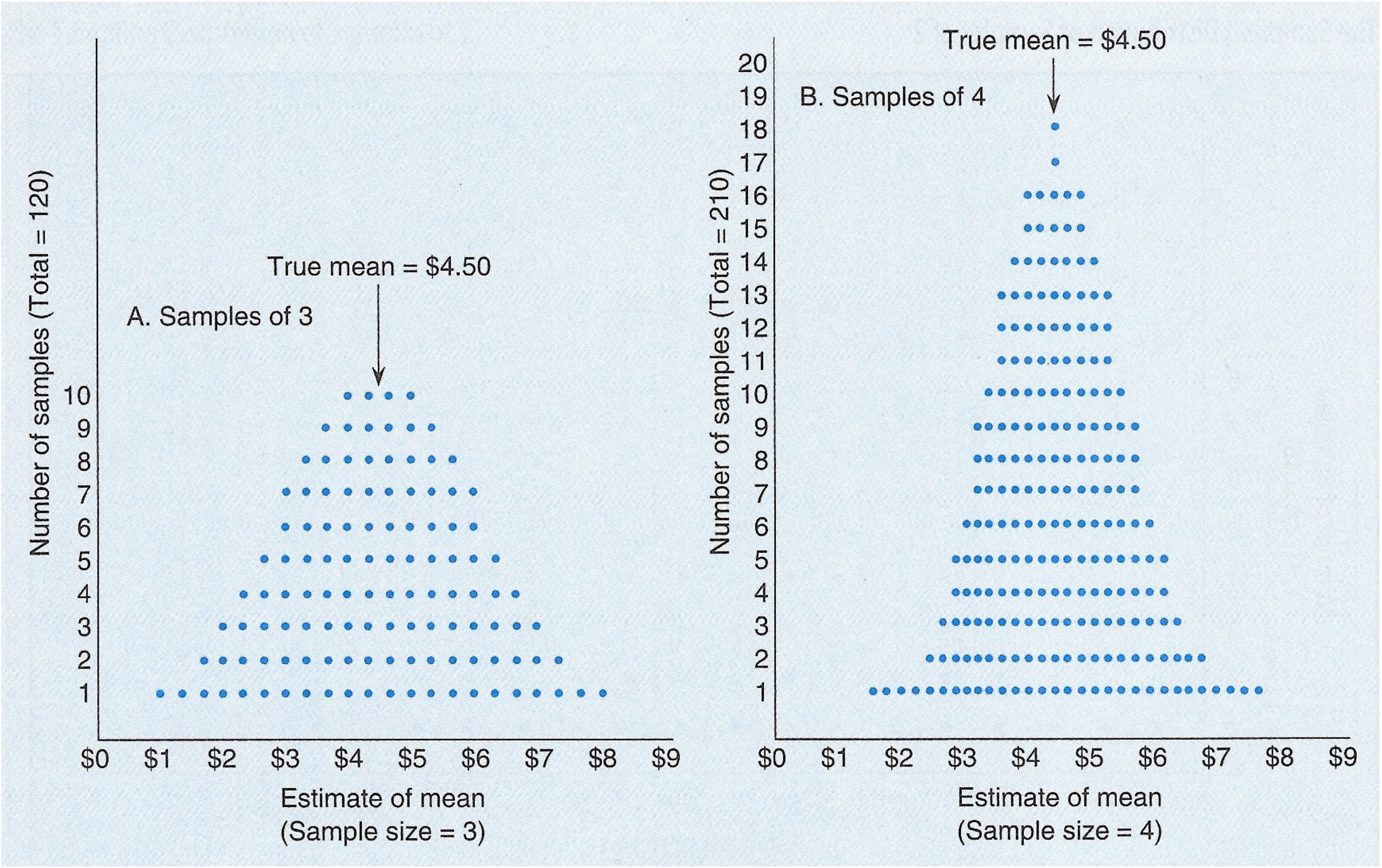
Source: Babbie 2001, p.188.

The sampling distribution (*n*=2)



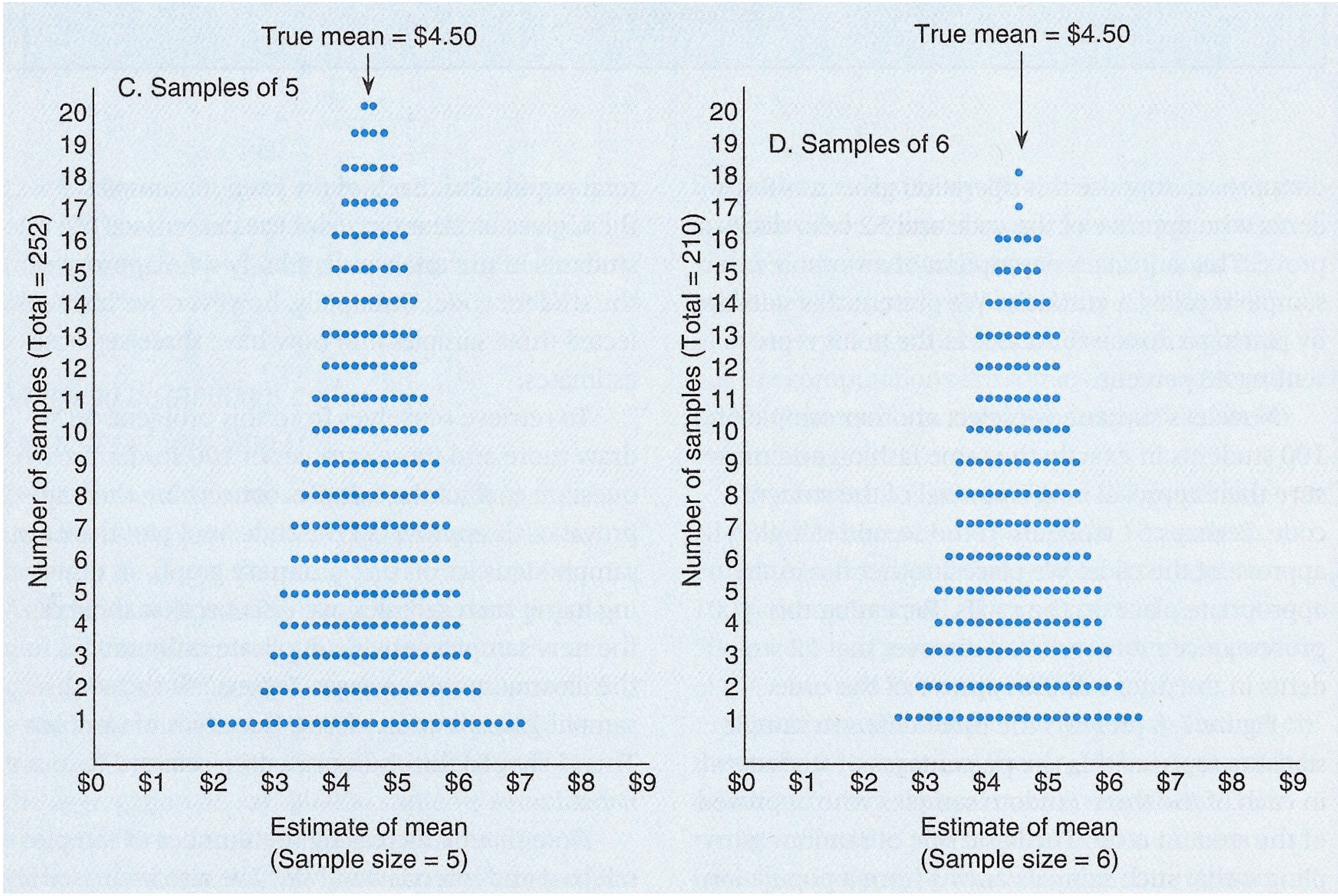
Source: Babbie 2001, p.189.

The sampling distribution



Source: Babbie 2001, p.190.

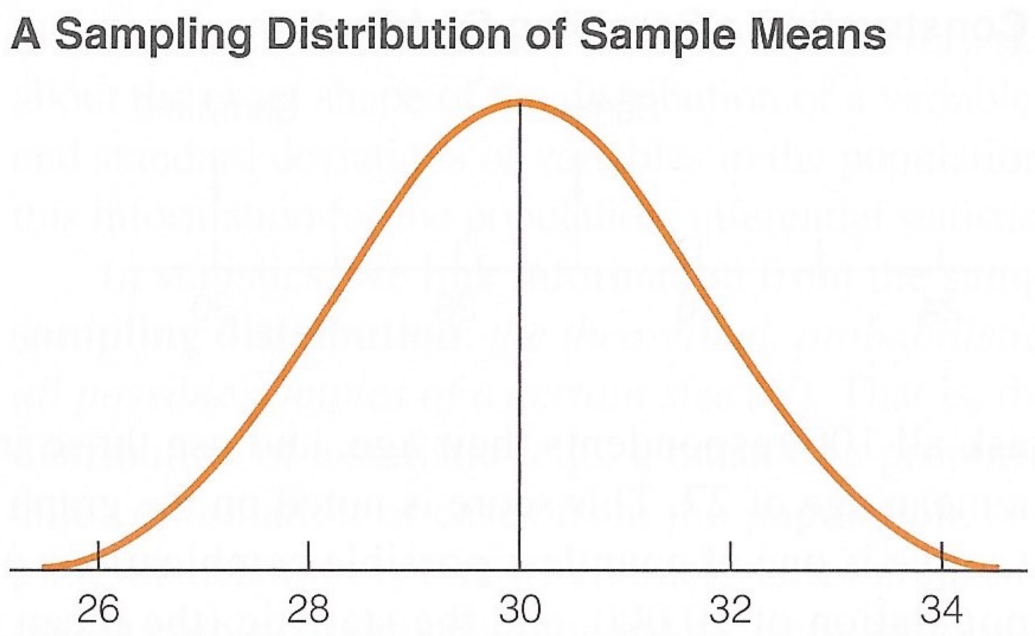
The sampling distribution



Source: Babbie 2001, p.190.

Properties of sampling distribution

- It has a mean ($\mu_{\bar{X}}$) equal to the population mean (μ)
- It has a standard deviation (standard error, $\sigma_{\bar{X}}$) equal to the population standard deviation (σ) divided by the square root of *n*
- It has a normal distribution



Source: Healey 2015, p.152.

First theorem

- Tells us the shape of the sampling distribution and defines its mean and standard deviation

- If repeated random samples of size *n* are drawn from a **normal population** with mean *μ* and standard deviation *σ*
  - Then, the sampling distribution of sample means will **have a normal distribution** with...
  - A mean: $\mu_{\bar{X}} = \mu$
  - A standard error of the mean: $\sigma_{\bar{X}} = \sigma/\sqrt{n}$

First theorem

- Begin with a characteristic that is normally distributed across a population (IQ, height)
- Take an infinite number of equally sized random samples from that population
- The sampling distribution of sample means will be normal

Central limit theorem

- If repeated random samples of size *n* are drawn from **any population** with mean μ and standard deviation σ
  - Then, as *n* becomes large, the sampling distribution of sample means will **approach normality** with...
  - A mean: $\mu_{\bar{X}} = \mu$
  - A standard error of the mean: $\sigma_{\bar{X}} = \sigma/\sqrt{n}$
- This is true for any variable, even those that are not normally distributed in the population
  - As sample size grows larger, the sampling distribution of sample means will become normal in shape

Central limit theorem

- The importance of the central limit theorem is that it removes the constraint of normality in the population
  - Applies to large samples ($n \geq 100$)

- If the sample is small ($n < 100$)
  - We must have information on the normality of the population before we can assume the sampling distribution is normal

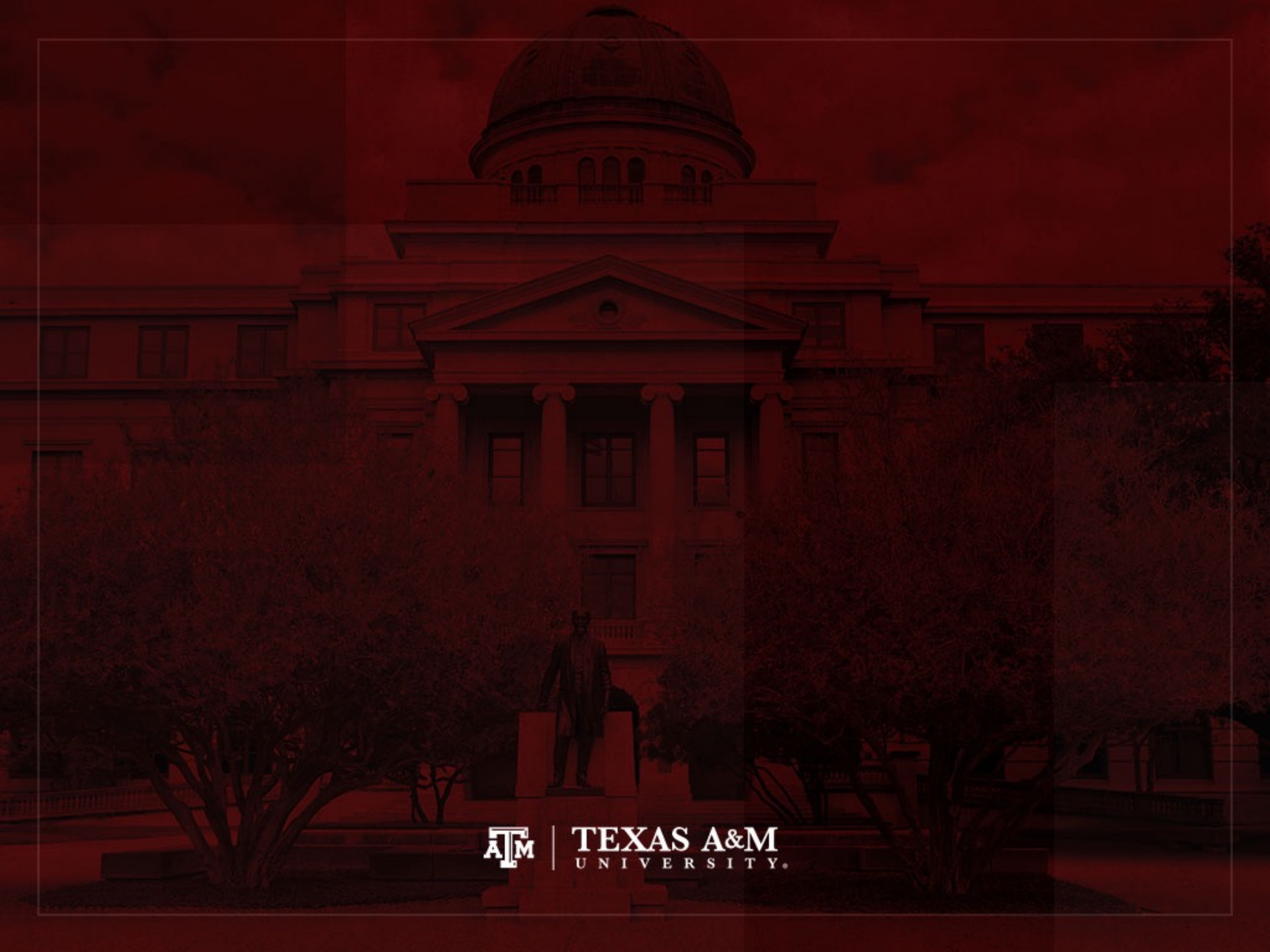
Additional considerations

- The sampling distribution is normal
  - We can estimate areas under the curve (Appendix A)
  - Or in Stata: **`display normal(z)`**
- We do not know the value of the population mean (μ)
  - But the mean of the sampling distribution ($\mu_{\bar{X}}$) is the same value as μ
- We do not know the value of the population standard deviation (σ)
  - But the standard deviation of the sampling distribution ($\sigma_{\bar{X}}$) is equal to σ divided by the square root of n

Symbols

Distribution	Shape	Mean	Standard deviation	Proportion
Samples	Varies	$\bar{X}$	s	P_s
Populations	Varies	μ	σ	P_u
Sampling distributions	Normal	$\mu_{\bar{X}}$		
of means		$\mu_{\bar{X}}$	$\sigma_{\bar{X}} = \sigma/\sqrt{n}$	
of proportions		μ_p	σ_p	

Source: Healey 2015, p.157.


TEXAS A&M
UNIVERSITY

Estimation procedures

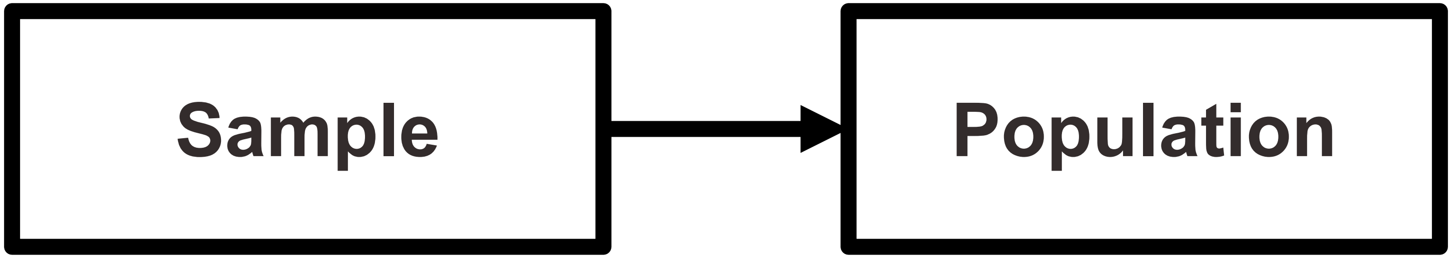
- Explain the logic of estimation, role of the sample, sampling distribution, and population
- Define and explain the concepts of bias and efficiency
- Construct and interpret confidence intervals for sample means and sample proportions
- Explain relationships among confidence level, sample size, and width of the confidence interval

Sample and population

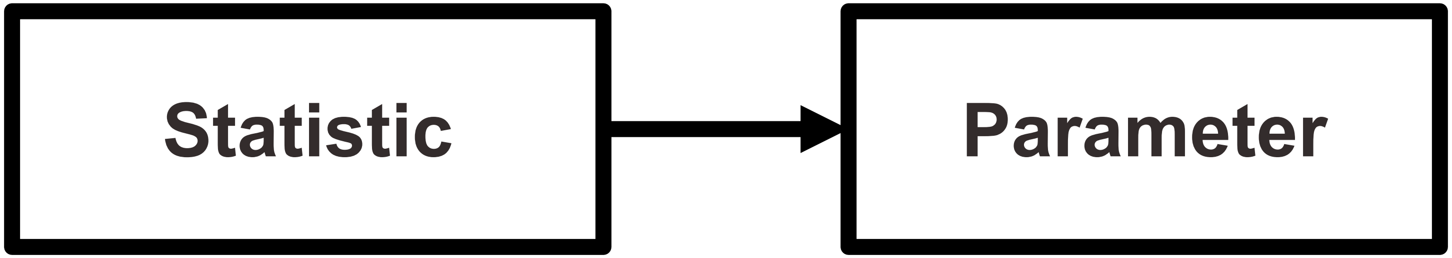
- In estimation procedures, statistics calculated from random samples are used to estimate the value of population parameters
- Example
  - If we know that 42% of a random sample drawn from a city are Republicans, we can estimate the percentage of all city residents who are Republicans

Terminology

- Information from samples is used to estimate information about the population



- Statistics are used to estimate parameters



Basic logic

- Sampling distribution is the link between sample and population

- The values of the parameters are unknown, but the characteristics of the sampling distribution are defined by two theorems (previous chapter)

Two estimation procedures

- **A point estimate** is a sample statistic used to estimate a population value
  - 68% of a sample of randomly selected Americans support capital punishment (GSS 2010)
- **An interval estimate** consists of confidence intervals (range of values)
  - Between 65% and 71% of Americans approve of capital punishment (GSS 2010)
  - Most point estimates are actually interval estimates
  - Margin of error generates confidence intervals
  - Estimators are selected based on two criteria
    - Bias (mean) and efficiency (standard error)

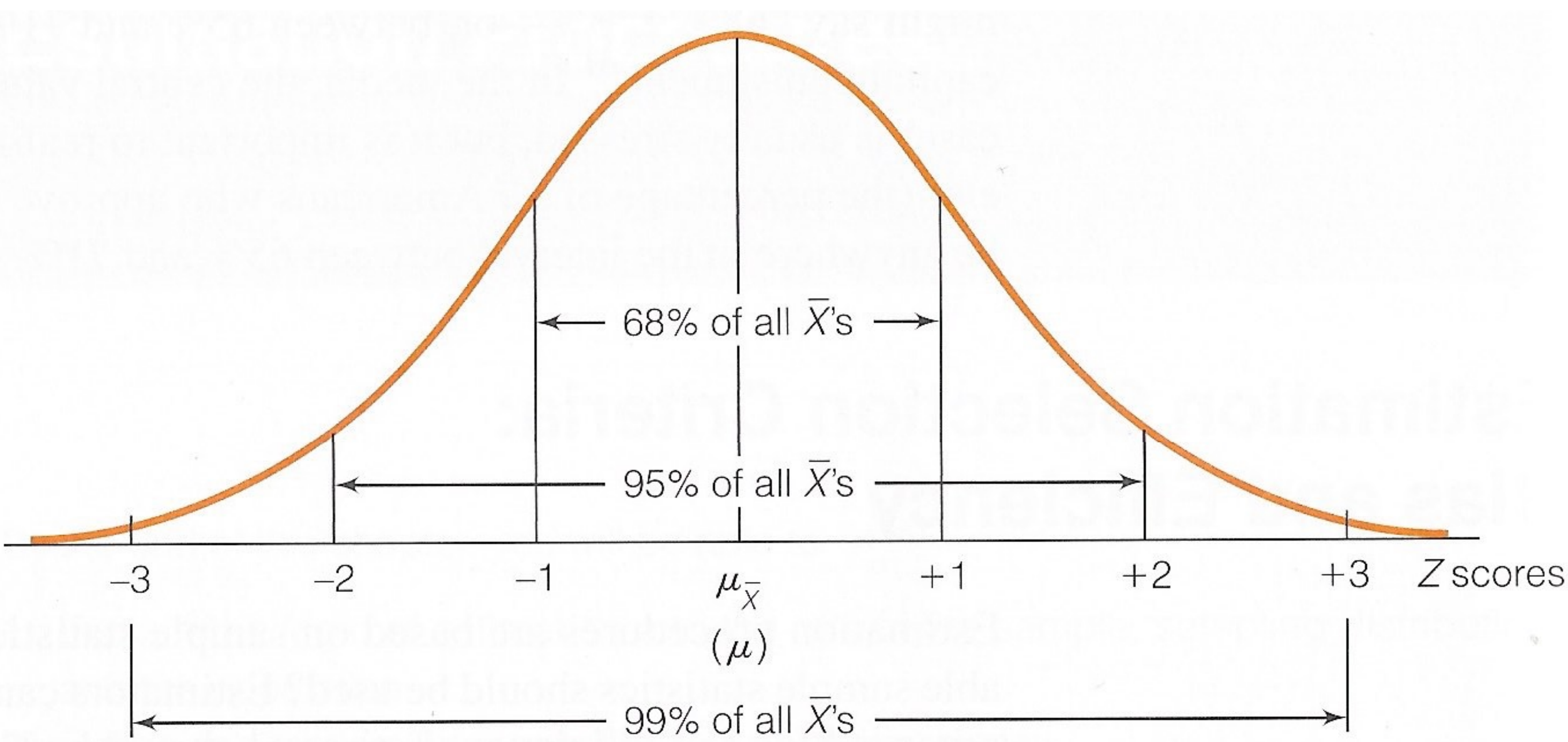
Bias

- An estimator is unbiased if the mean of its sampling distribution is equal to the population value of interest
- The mean of the sampling distribution of sample means ($\mu_{\bar{X}}$) is the same as the population mean (μ)
- Sample proportions (P_s) are also unbiased
  - If we calculate sample proportions from repeated random samples of size *n*...
  - Then, the sampling distribution of sample proportions will have a mean (μ_p) equal to the population proportion (P_u)
- Sample means and proportions are unbiased estimators
  - We can determine the probability that they are within a certain distance of the population values

Example

- Random sample to get income information
- Sample size (*n*): 500 households
- Sample mean ($\bar{X}$): $45,000
- Population mean (*μ*): unknown parameter
- Mean of sampling distribution ($\mu_{\bar{X}} = \mu$)
  - If an estimator ($\bar{X}$) is unbiased, it is probably an accurate estimate of the population parameter (*μ*) and sampling distribution mean ($\mu_{\bar{X}}$)
  - We use the sampling distribution (which has a normal shape) to estimate confidence intervals

Sampling distribution



Source: Healey 2015, p.162.

Efficiency

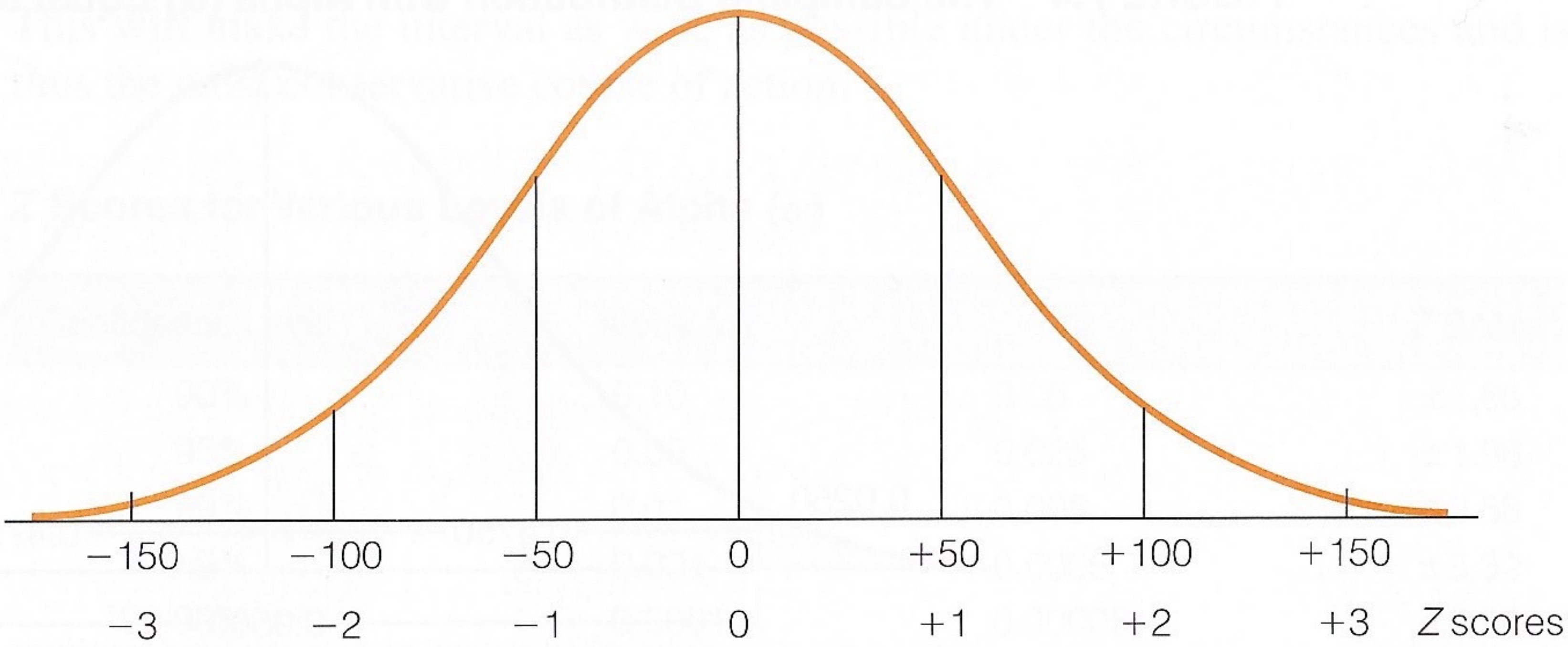
- Efficiency is the extent to which the sampling distribution is clustered around its mean
- Efficiency or clustering is a matter of dispersion
  - The smaller the standard deviation of a sampling distribution, the greater the clustering and the higher the efficiency
  - Larger samples have greater clustering and higher efficiency
  - Standard deviation of sampling distribution: $\sigma_{\bar{X}} = \sigma/\sqrt{n}$

Statistics	Sample 1	Sample 2
Sample mean	$\bar{X}_1 = \$45{,}000$	$\bar{X}_2 = \$45{,}000$
Sample size	$n_1 = 100$	$n_2 = 1000$
Standard deviation	$\sigma_1 = \$500$	$\sigma_2 = \$500$
Standard error	$\sigma_{\bar{X}} = 500/\sqrt{100} = \$50.00$	$\sigma_{\bar{X}} = 500/\sqrt{1000} = \$15.81$

Source: Healey 2015, p.163.

Sampling distribution

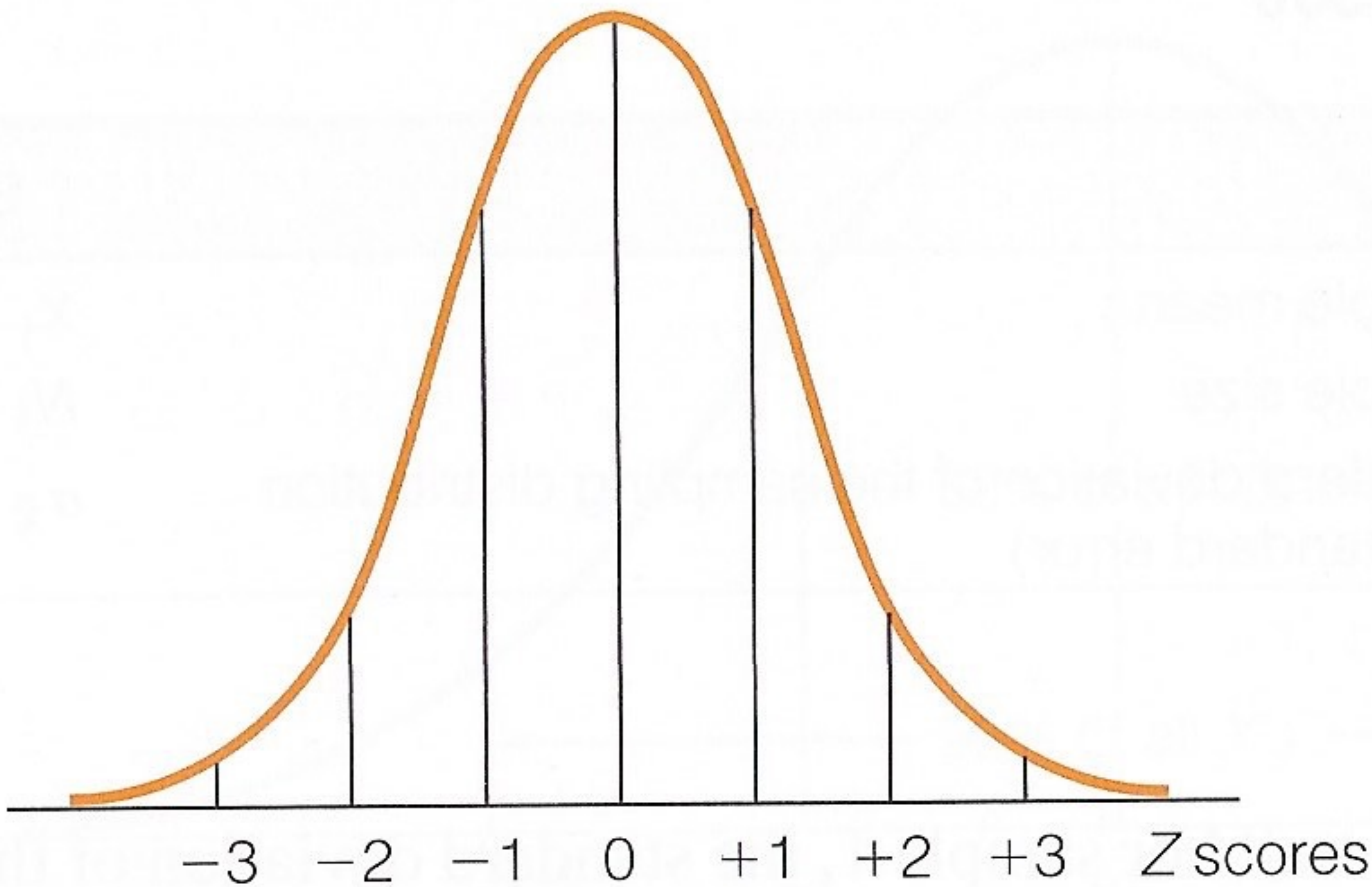
$n = 100; \sigma_{\bar{X}} = \$50.00$



Source: Healey 2015, p.163.

Sampling distribution

n = 1000; $\sigma_{\bar{X}}$ = \$15.81



Source: Healey 2015, p.164.

Confidence interval & level

- **Confidence interval** is a range of values used to estimate the true population parameter
  - We associate a confidence level (e.g. 0.95 or 95%) to a confidence interval
- **Confidence level** is the success rate of the procedure to estimate the confidence interval
  - Expressed as probability ($1–\alpha$) or percentage $(1–\alpha)*100$
  - α is the complement of the confidence level
  - Larger confidence levels generate larger confidence intervals
- Confidence level of 95% is the most common
  - Good balance between precision (width of confidence interval) and reliability (confidence level)

Interval estimation procedures

- Set the alpha (α)
  - Probability that the interval will be wrong
- Find the *Z* score associated with alpha
  - In column c of Appendix A of textbook
    - If the *Z* score you are seeking is between two other scores, choose the larger of the two *Z* scores
  - In Stata: **`display invnormal(`α`)`**
- Substitute values into appropriate equation
- Interpret the interval

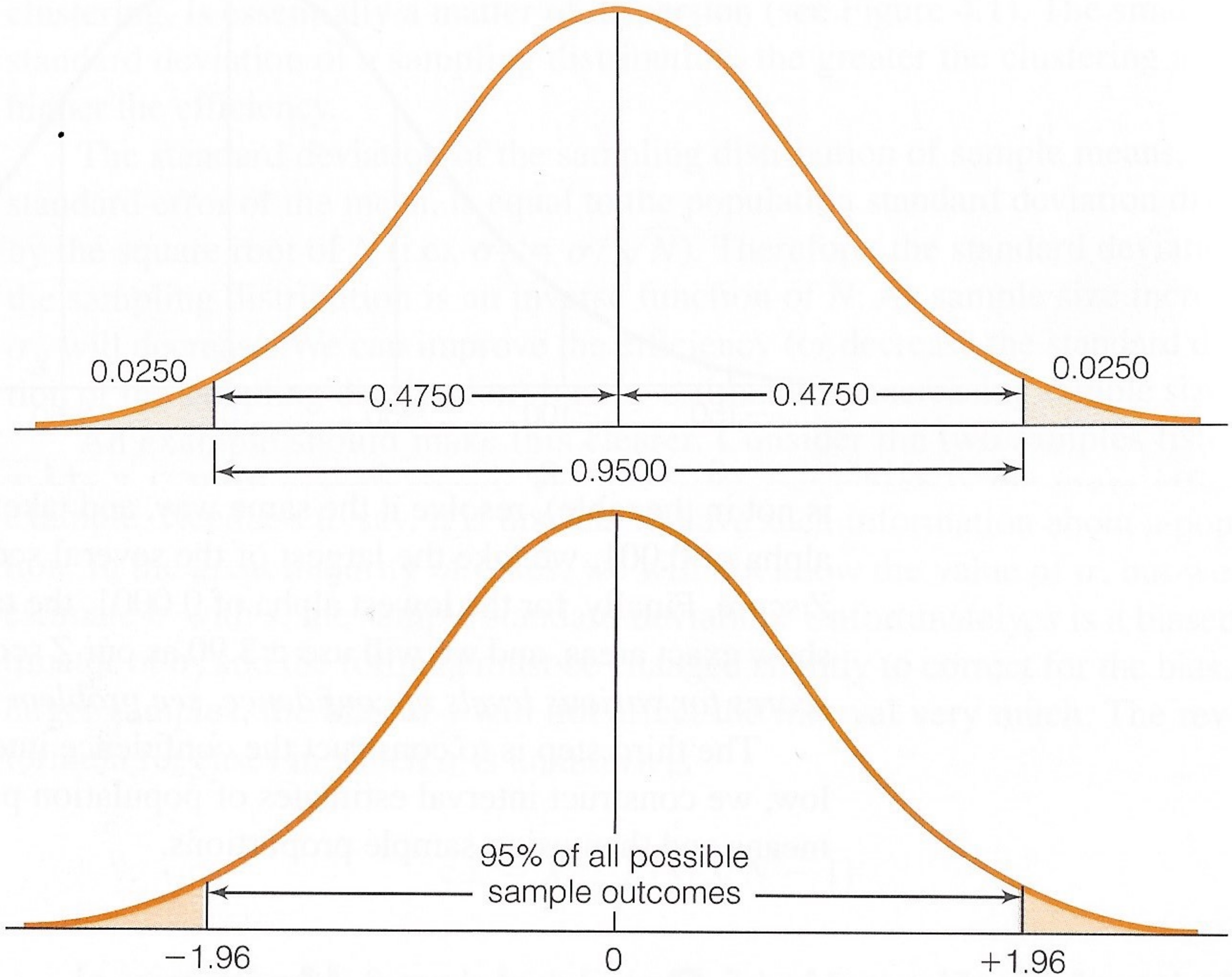
Example to find *Z* score

- Setting alpha (*α*) equal to 0.05
  - 95% confidence level: (1–*α*)*100
  - We are willing to be wrong 5% of the time
- If alpha is equal to 0.05
  - Half of this probability is in the lower tail (*α*/2=0.025)
  - Half is in the upper tail of the distribution (*α*/2=0.025)
- Looking up this area, we find a Z = 1.96

```
di invnormal(.025)
-1.959964
```

```
di invnormal(1-.025)
di invnormal(.975)
1.959964
```

Finding Z for sampling distribution with $\alpha = 0.05$



Source: Healey 2015, p.164–165.

Confidence level, α, and Z

Confidence level $(1 - \alpha) * 100$	Significance level alpha (α)	$\alpha / 2$	Z score
90%	0.10	0.05	$\pm$1.65
95%	**0.05**	**0.025**	**$\pm$1.96**
99%	0.01	0.005	$\pm$2.58
99.9%	0.001	0.0005	$\pm$3.32
99.99%	0.0001	0.00005	$\pm$3.90

Source: Healey 2015, p.165.

Confidence intervals for sample means

- For large samples ($n \geq 100$)
- Standard deviation (σ) **known** for population

$$c.i. = \bar{X} \pm Z\left(\frac{\sigma}{\sqrt{n}}\right)$$

c.i. = confidence interval

$\bar{X}$ = sample mean

Z = score determined by the alpha level (confidence level)

$\sigma/\sqrt{n}$ = sample deviation of the sampling distribution (standard error of the mean)

$\pm Z(\sigma/\sqrt{n})$ = margin of error

Example for means: Large sample, σ known

- Sample of 200 residents
- Sample mean of IQ is 105
- Population standard deviation is 15
- Calculate a confidence interval with a 95% confidence level (α = 0.05)
  - Same as saying: calculate a 95% confidence interval

$$c.i. = \bar{X} \pm Z\left(\frac{\sigma}{\sqrt{n}}\right) = 105 \pm 1.96\left(\frac{15}{\sqrt{200}}\right) = 105 \pm 2.08$$

  - Average IQ is somewhere between 102.92 (105–2.08) and 107.08 (105+20.8)

Interpreting previous example

$$n = 200; 102.92 \le \mu \le 107.08$$

- **Correct:** We are 95% certain that the confidence interval contains the true value of μ
  - If we selected several samples of size 200 and estimated their confidence intervals, 95% of them would contain the population mean (μ)
  - The 95% confidence level refers to the success rate to estimate the population mean (μ). It does not refer to the population mean itself
- **Wrong:** Since the value of μ is fixed, it is incorrect to say that there is a chance of 95% that the true value of μ is between the interval

Source: Triola 2008, p.250–303.

Confidence intervals for sample means

- For large samples ($n \geq 100$)
- Standard deviation (σ) **unknown** for population

$$c.i. = \bar{X} \pm Z\left(\frac{s}{\sqrt{n-1}}\right)$$

c.i. = confidence interval

$\bar{X}$ = sample mean

Z = score determined by the alpha level (confidence level)

$s/\sqrt{n-1}$ = sample deviation of the sampling distribution (standard error of the mean)

$\pm Z(s/\sqrt{n-1})$ = margin of error

Example for means: Large sample, σ unknown

- Sample of 500 residents
- Sample mean income is \$45,000
- Sample standard deviation is \$200
- Calculate a 95% confidence interval

$$c.i. = \bar{X} \pm Z\left(\frac{s}{\sqrt{n-1}}\right) = 45{,}000 \pm 1.96\left(\frac{200}{\sqrt{500-1}}\right)$$

$$c.i. = 45{,}000 \pm 17.54$$

  – Average income is between \$44,982.46 (45,000–17.54) and \$45,017.54 (45,000+17.54)

Example from ACS

- We are 95% certain that the confidence interval from $49,926.89 to $50,161.07 contains the true average wage and salary income for the U.S. population in 2018

```
. ***95% confidence level
. svy, subpop(if income!=. & income!=0): mean income
(running mean on estimation sample)

Survey: Mean estimation

Number of strata =   2,351      Number of obs     =     3,214,539
Number of PSUs   = 1410976      Population size   =   327,167,439
                                Subpop. no. obs   =     1,574,313
                                Subpop. size      =   163,349,075
                                Design df         =     1,408,625
```

	Mean	Linearized Std. Err.	[95% Conf.	Interval]
income	50043.98	59.74195	49926.89	50161.07

```
.
. ***Standard deviation
. estat sd
```

	Mean	Std. Dev.
income	50043.98	61547.67

Obs.: Only individuals with some wage and salary income are included (exclude those with zero income).

Source: 2018 American Community Survey.

Edited table

Table 1. Summary statistics for individual average wage and salary income of the U.S. population, 2018

Summary statistics	Value
Mean	50,043.98
Standard deviation	61,547.67
Standard error	59.74
95% confidence interval	
Lower bound	49,926.89
Upper bound	50,161.07
Sample size	1,574,313

Obs.: Only individuals with some wage and salary income are included (exclude those with zero income).
Source: 2018 American Community Survey.

Interpreting previous example

n = 1,574,313; 49,926.89 $\leq \mu \leq$ 50,161.07

- **Correct:** We are 95% certain that the confidence interval contains the true value of μ
  - If we selected several samples of size 1,574,313 and estimated their confidence intervals, 95% of them would contain the population mean (μ)
  - The 95% confidence level refers to the success rate to estimate the population mean (μ). It does not refer to the population mean itself
- **Wrong:** Since the value of μ is fixed, it is incorrect to say that there is a chance of 95% that the true value of μ is between the interval

Source: Triola 2008, p.250–303.

Example from GSS

- We are 95% certain that the confidence interval from \$35,324.83 to \$39,889.96 contains the true average income for the U.S. adult population in 2004

```
. svy: mean conrinc, over(year)
(running mean on estimation sample)

Survey: Mean estimation

Number of strata =     307          Number of obs   =      4,522
Number of PSUs   =     597          Population size = 4,611.7099
                                    Design df       =        290

         2004: year = 2004
         2010: year = 2010
         2016: year = 2016
```

Over	Mean	Linearized Std. Err.	[95% Conf.	Interval]
conrinc				
2004	37607.39	1159.734	35324.83	39889.96
2010	31537.11	1216.566	29142.69	33931.53
2016	34649.3	1267.614	32154.41	37144.19

Note: Variance scaled to handle strata with a single sampling unit.

Source: 2004, 2010, 2016 General Social Surveys.

Edited table

Table 1. Mean, standard error, 95% confidence interval, and sample size of individual average income of the U.S. adult population, 2004, 2010, and 2016

Year	Mean	Standard Error	95% Confidence Interval		Sample Size
			Lower Bound	Upper Bound	
2004	37,607.39	1,159.73	35,324.83	39,889.96	1,688
2010	31,537.11	1,216.57	29,142.69	33,931.53	1,202
2016	34,649.30	1,267.61	32,154.41	37,144.19	1,632

Source: 2004, 2010, 2016 General Social Surveys.

Confidence intervals for sample proportions

$$c.i. = P_s \pm Z\sqrt{\frac{P_u(1-P_u)}{n}}$$

c.i. = confidence interval

P_s = sample proportion

Z = score determined by the alpha level (confidence level)

$\sqrt{P_u(1-P_u)/n}$ = sample deviation of the sampling distribution (standard error of the proportion)

$\pm Z(\sqrt{P_u(1-P_u)/n})$ = margin of error

Note about sample proportions

- The formula for the standard error includes the population value
  - We do not know and are trying to estimate (P_u)
- By convention we set P_u equal to 0.50
  - The numerator [$P_u(1–P_u)$] is at its maximum value
  - $P_u(1–P_u) = (0.50)(1–0.50) = 0.25$
- The calculated confidence interval will be at its maximum width
  - This is considered the most statistically conservative technique

Example for proportions

- Estimate the proportion of students who missed at least one day of classes last semester
  - In a random sample of 200 students, 60 students reported missing one day of class last semester
  - Thus, the sample proportion is 0.30 (60/200)
  - Calculate a 95% (alpha = 0.05) confidence interval

$$c.i. = P_s \pm Z\sqrt{\frac{P_u(1-P_u)}{n}} = 0.3 \pm 1.96\sqrt{\frac{0.5(1-0.5)}{200}}$$

$$c.i. = 0.3 \pm 0.08$$

Example from ACS

- We are 95% certain that the confidence interval from 5.2% to 5.3% contains the true proportion of internal migrants in the U.S. population in 2018

```
. svy: prop migrant
(running proportion on estimation sample)

Survey: Proportion estimation

Number of strata =    2,351                 Number of obs     =     3,184,099
Number of PSUs   = 1410889                  Population size =   323,541,502
                                            Design df         =     1,408,538
```

	Proportion	Linearized Std. Err.	Logit [95% Conf.	Interval]
migrant				
Non-migrant	.9418963	.000259	.9413866	.9424019
Internal migrant	.0524799	.0002463	.0519993	.0529647
International migrant	.0056239	.0000823	.0054649	.0057874

Source: 2018 American Community Survey.

Edited table

Table 2. Summary statistics for migration status of the U.S. population, 2018

Migration status	Proportion	Standard Error	95% Confidence Interval	
			Lower Bound	Upper Bound
Non-migrant	0.9419	0.0003	0.9414	0.9424
Internal migrant	0.0525	0.0003	0.0520	0.0530
International migrant	0.0056	0.0001	0.0055	0.0058

Obs.: Sample size of 3,184,099 individuals.

Source: 2018 American Community Survey.

Interpreting previous example

n = 3,184,099; $5.2 \leq P_u \leq 5.3$

- **Correct:** We are 95% certain that the confidence interval contains the true value of P_u
  - If we selected several samples of size 3,184,099 and estimated their confidence intervals, 95% of them would contain the population proportion (P_u)
  - The 95% confidence level refers to the success rate to estimate the population proportion (P_u). It does not refer to the population proportion itself
- **Wrong:** Since the value of P_u is fixed, it is incorrect to say that there is a chance of 95% that the true value of P_u is between the interval

Source: Triola 2008, p.250–303.

Example from GSS

- We are 95% certain that the confidence interval from 2.6% to 4.7% contains the true proportion of the U.S. adult population who thinks the number of immigrants to the country should increase a lot in 2004

```
. svy: prop letin1 if year==2004
(running proportion on estimation sample)

Survey: Proportion estimation

Number of strata =     109          Number of obs   =      1,983
Number of PSUs   =     218          Population size = 1,979.3435
                                    Design df       =        109

      _prop_1: letin1 = increased a lot
      _prop_2: letin1 = increased a little
      _prop_3: letin1 = remain the same as it is
      _prop_4: letin1 = reduced a little
      _prop_5: letin1 = reduced a lot
```

	Proportion	Linearized Std. Err.	[95% Conf.	Interval]
letin1				
_prop_1	.0348265	.005221	.0258369	.0467936
_prop_2	.0653852	.0060495	.0543699	.078447
_prop_3	.3517117	.0128957	.3265967	.3776749
_prop_4	.2829629	.0118188	.2601357	.3069621
_prop_5	.2651137	.0127052	.2407073	.2910462

Source: 2004 General Social Survey.

Edited table

Table 2. Proportion, standard error, 95% confidence interval, and sample size of opinion of the U.S. adult population about how should the number of immigrants to the country be nowadays, 2004, 2010, and 2016

Opinion About Number of Immigrants	Proportion	Standard Error	95% Confidence Interval		Sample Size
			Lower Bound	Upper Bound	
2004					1,983
Increase a lot	0.0348	0.0052	0.0258	0.0468	
Increase a little	0.0654	0.0060	0.0544	0.0784	
Remain the same	0.3517	0.0129	0.3266	0.3777	
Reduce a little	0.2830	0.0118	0.2601	0.3070	
Reduce a lot	0.2651	0.0127	0.2407	0.2910	
2010					1,393
Increase a lot	0.0426	0.0061	0.0320	0.0564	
Increase a little	0.0944	0.0096	0.0771	0.1152	
Remain the same	0.3589	0.0166	0.3268	0.3923	
Reduce a little	0.2452	0.0121	0.2220	0.2700	
Reduce a lot	0.2588	0.0146	0.2310	0.2887	
2016					1,845
Increase a lot	0.0586	0.0069	0.0462	0.0740	
Increase a little	0.1163	0.0091	0.0993	0.1358	
Remain the same	0.4028	0.0117	0.3797	0.4264	
Reduce a little	0.2305	0.0097	0.2118	0.2504	
Reduce a lot	0.1918	0.0101	0.1724	0.2128	

Source: 2004, 2010, 2016 General Social Surveys.

Width of confidence interval

- The width of confidence intervals can be controlled by manipulating the confidence level
  - The confidence level increases
  - The alpha decreases
  - The *Z* score increases
  - The confidence interval is wider

Example: $\bar{X}$ = \$45,000; *s* = \$200; *n* = 500

Confidence level	Alpha (α)	*Z* score	Confidence interval	Interval width
90%	0.10	$\pm$1.65	\$45,000 $\pm$ \$14.77	\$29.54
95%	0.05	$\pm$1.96	\$45,000 $\pm$ \$17.54	\$35.08
99%	0.01	$\pm$2.58	\$45,000 $\pm$ \$23.09	\$46.18
99.9%	0.001	$\pm$3.32	\$45,000 $\pm$ \$29.71	\$59.42

Width of confidence interval

- The width of confidence intervals can be controlled by manipulating the sample size
  - The sample size increases
  - The confidence interval is narrower

Example: $\bar{X}$ = $45,000; *s* = $200; *α* = 0.05

n	Confidence interval	Interval width
100	*c.i.* = $45,000 ± 1.96(200/√99) = $45,000 ± $39.40	$78.80
500	*c.i.* = $45,000 ± 1.96(200/√499) = $45,000 ± $17.55	$35.10
1000	*c.i.* = $45,000 ± 1.96(200/√999) = $45,000 ± $12.40	$24.80
10000	*c.i.* = $45,000 ± 1.96(200/√9999) = $45,000 ± $3.92	$7.84

Summary: Confidence intervals

- Sample means, large samples (*n*>100), population standard deviation known

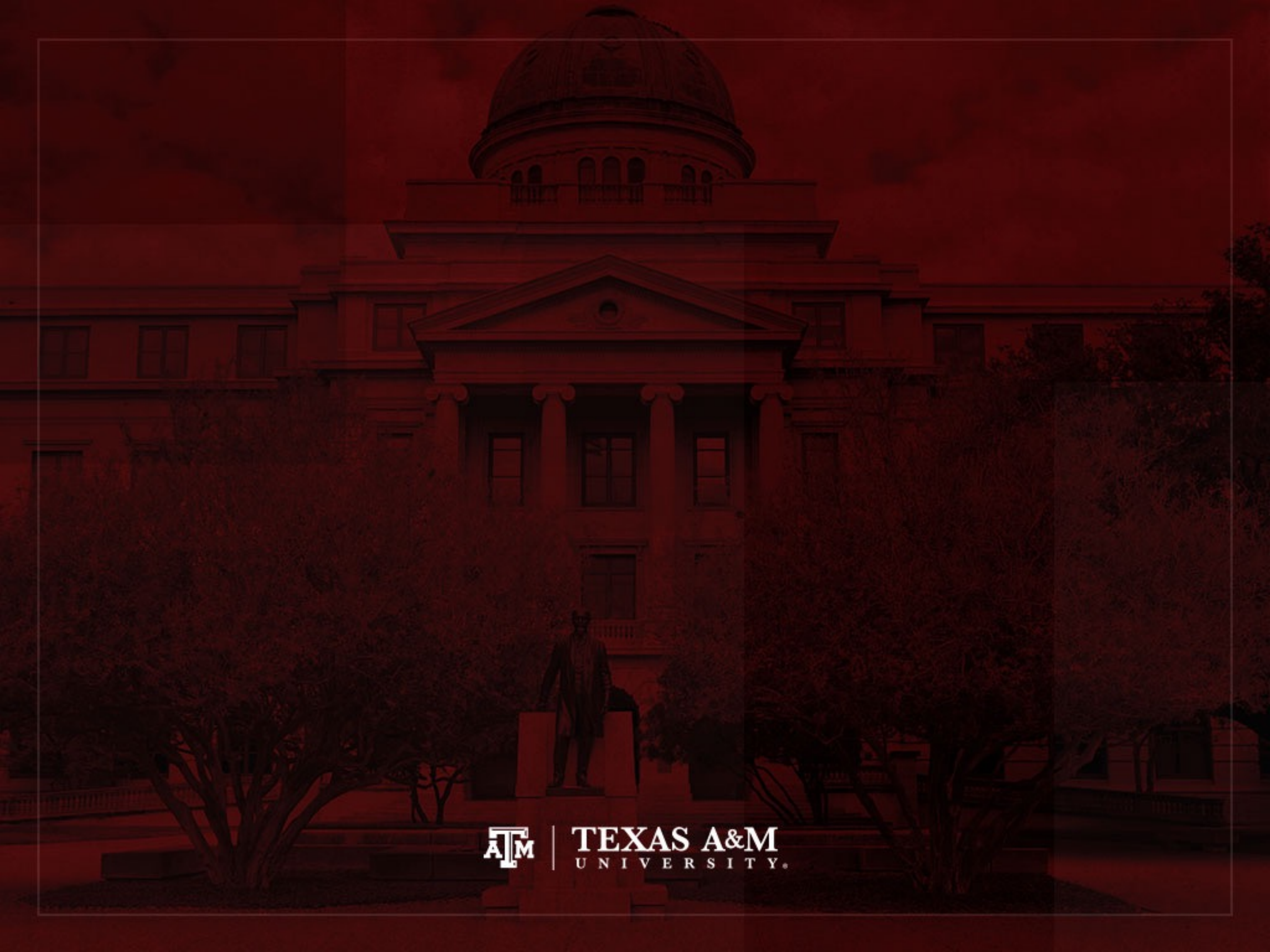
$$c.i. = \bar{X} \pm Z\left(\frac{\sigma}{\sqrt{n}}\right)$$

- Sample means, large samples (*n*>100), population standard deviation unknown

$$c.i. = \bar{X} \pm Z\left(\frac{s}{\sqrt{n-1}}\right)$$

- Sample proportions, large samples (*n*>100)

$$c.i. = P_s \pm Z\sqrt{\frac{P_u(1-P_u)}{n}}$$


TEXAS A&M
UNIVERSITY