Lecture 4: Normal curve

Ernesto F. L. Amaral

September 21, 2023
Introduction to Sociological Data Analysis (SOCI 600)

www.ernestoamaral.com

Source: Healey, Joseph F. 2015. "Statistics: A Tool for Social Research." Stamford: Cengage Learning. 10th edition. Chapter 5 (pp. 122–142).

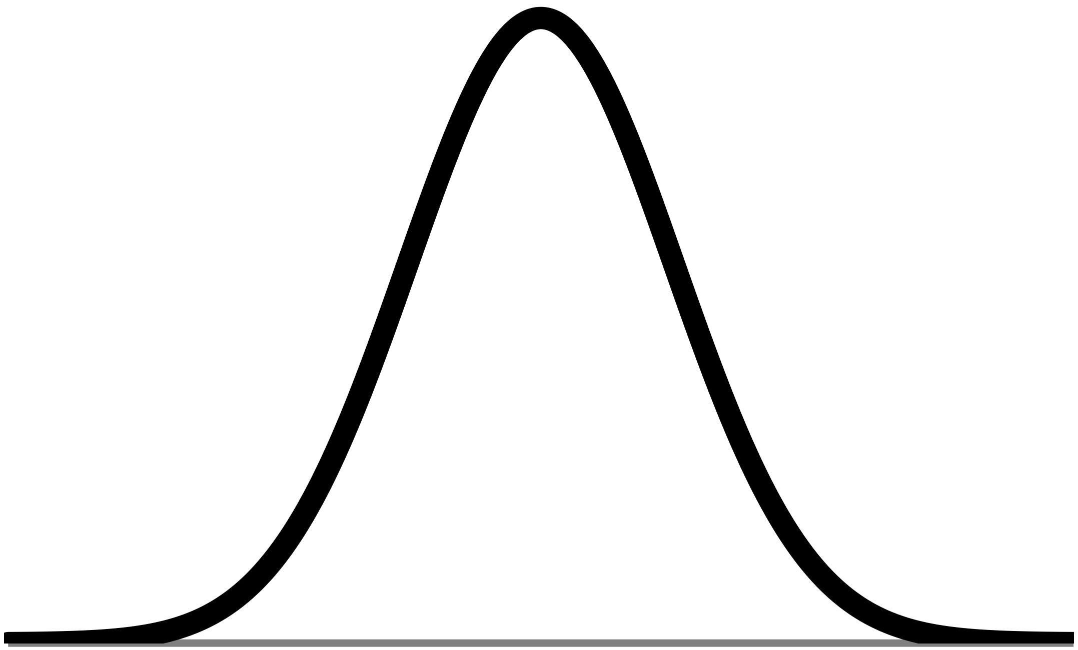
TEXAS A&M UNIVERSITY

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

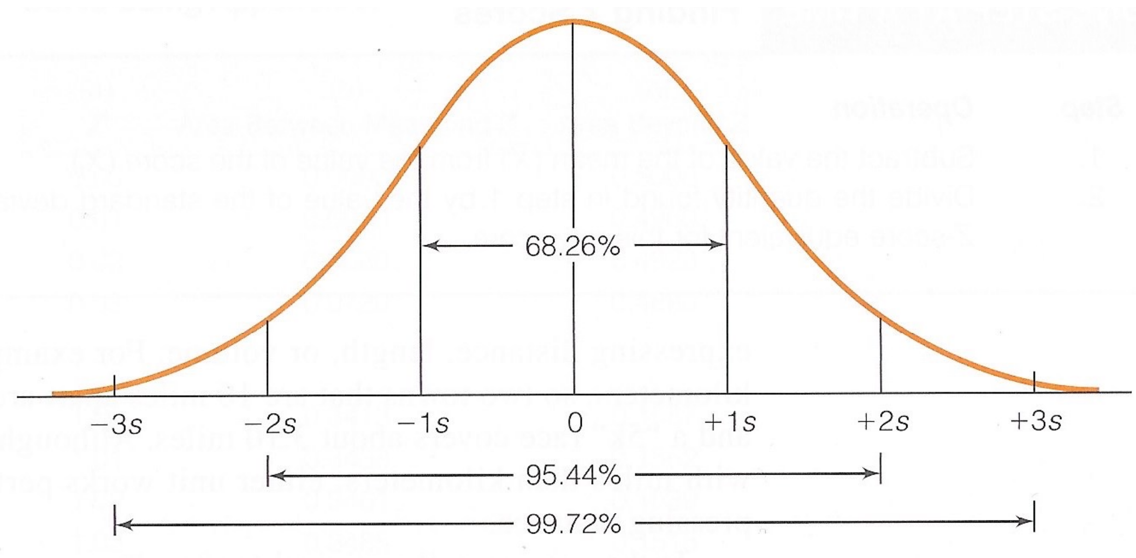
Properties of the normal curve

- 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%

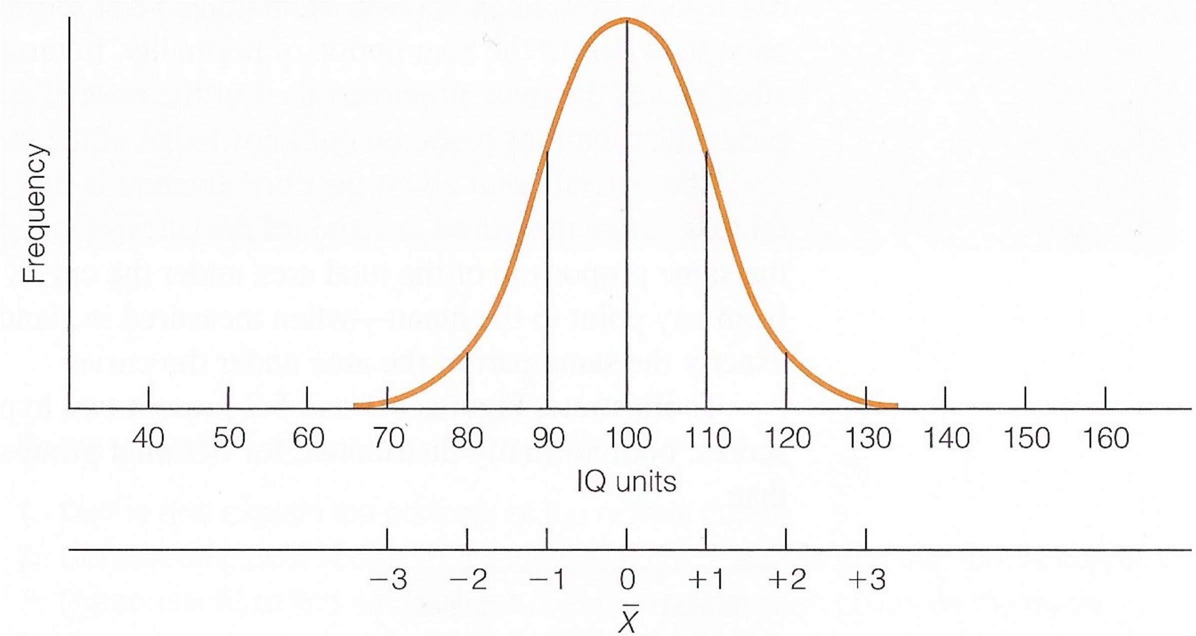
Source: Healey 2015, p.125.

IQ scores, females

$\bar{X}$ = 100

s = 10

N = 1000

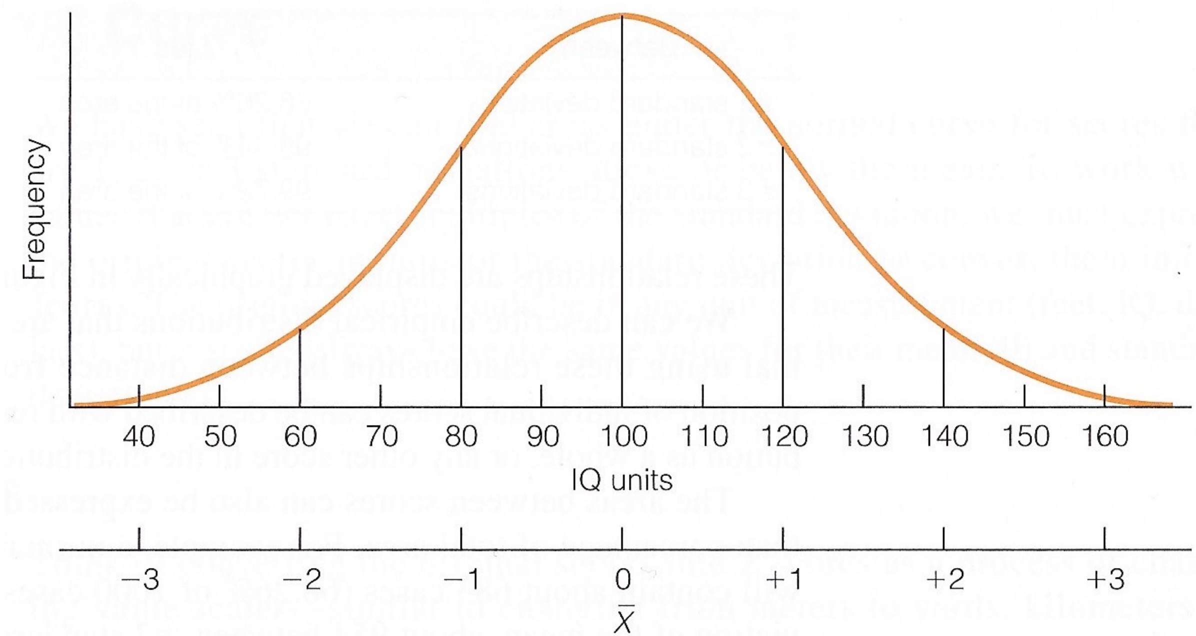


IQ scores, males

$\bar{X}$ = 100

s = 20

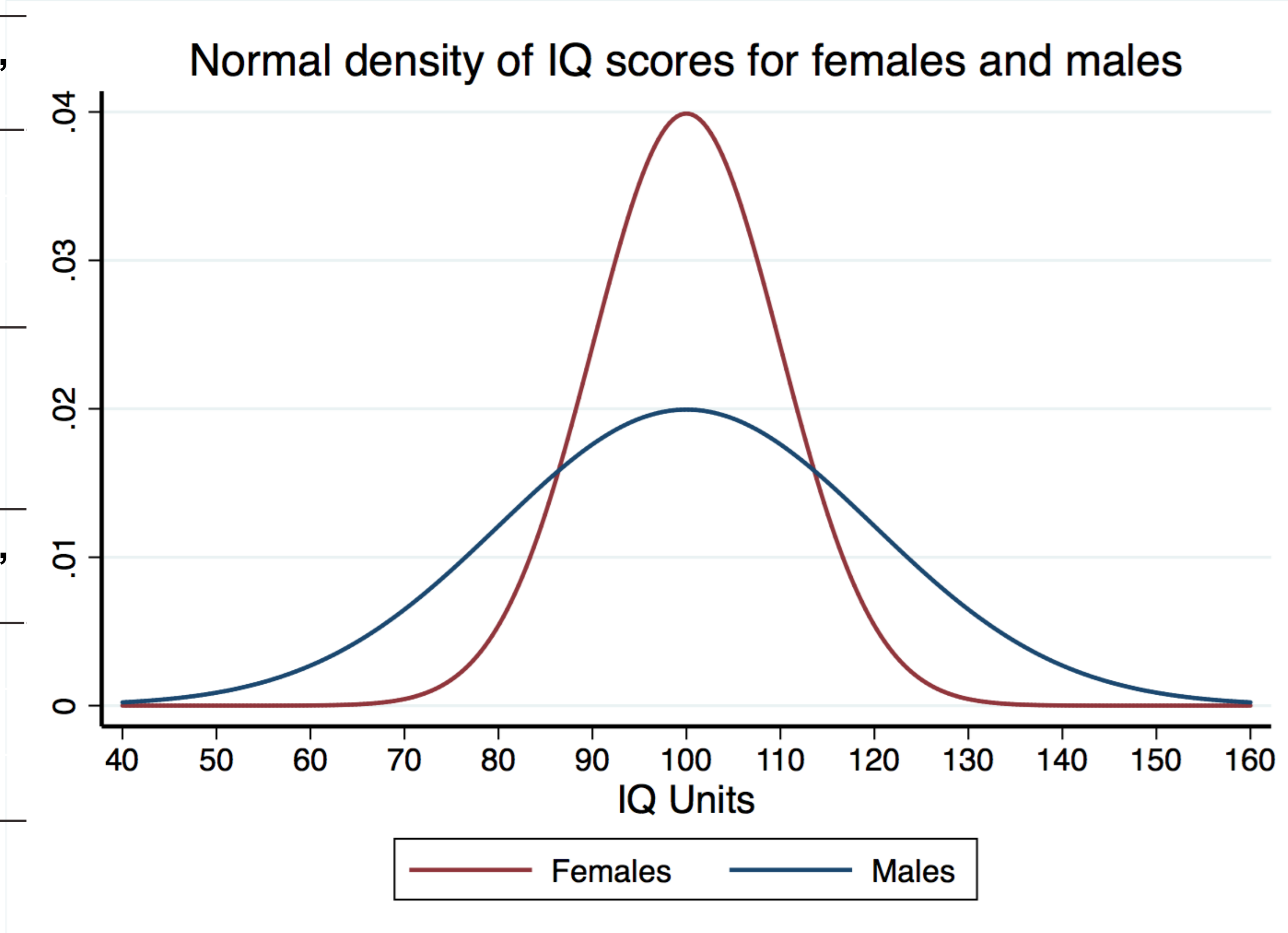
N = 1000



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.

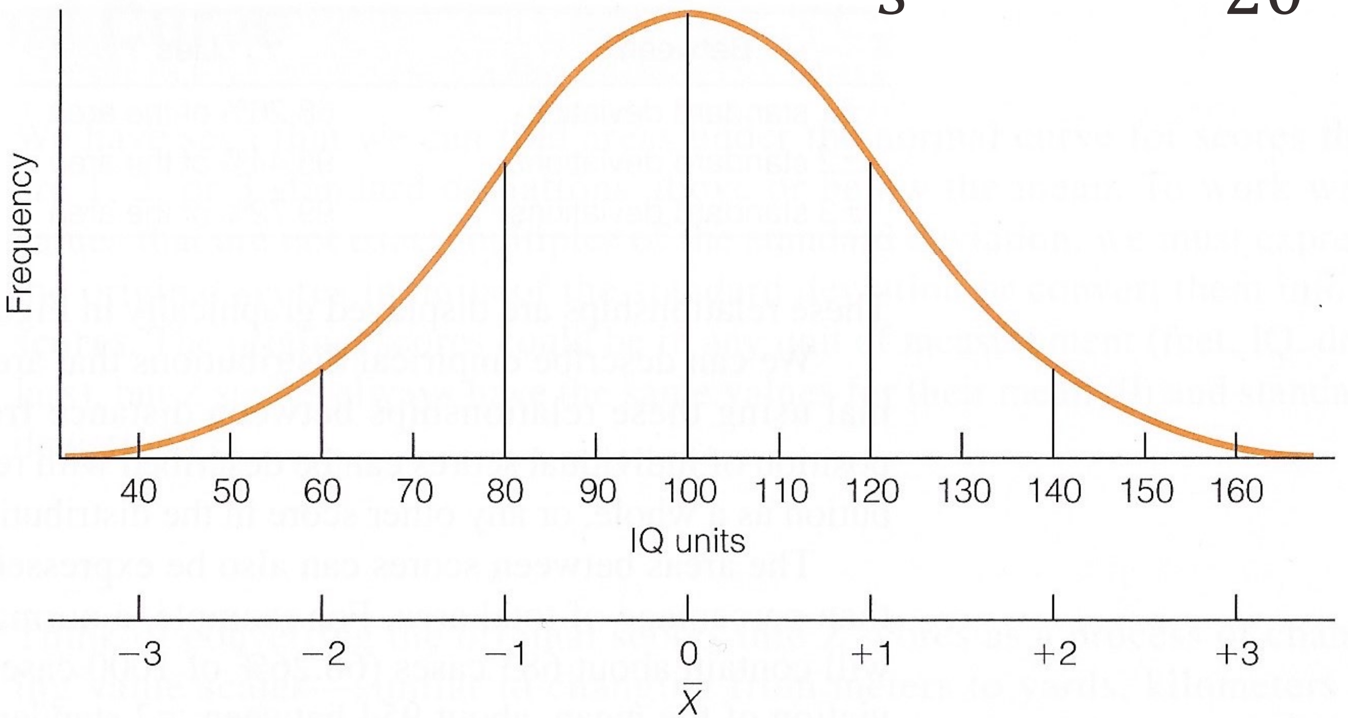
Z scores

- 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

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*

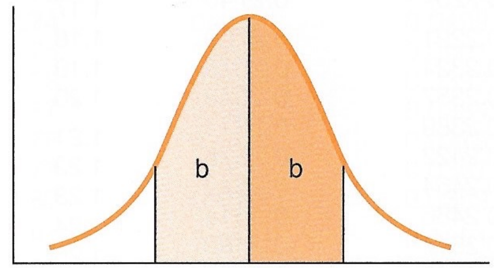
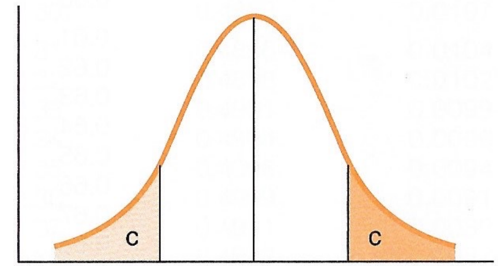


FIGURE A.2 Area Beyond *Z*



(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***

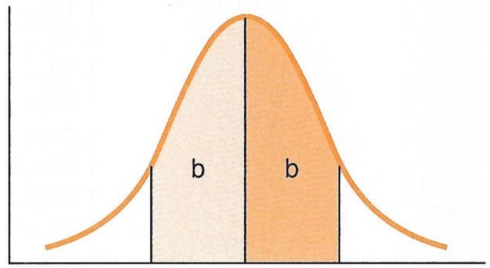
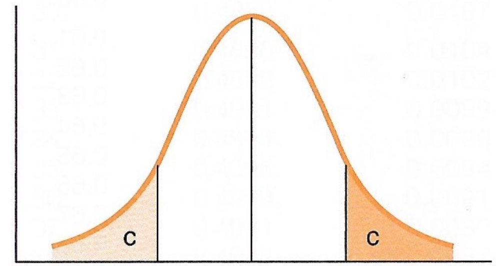


FIGURE A.2 **Area Beyond *Z***



(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
...	...	...

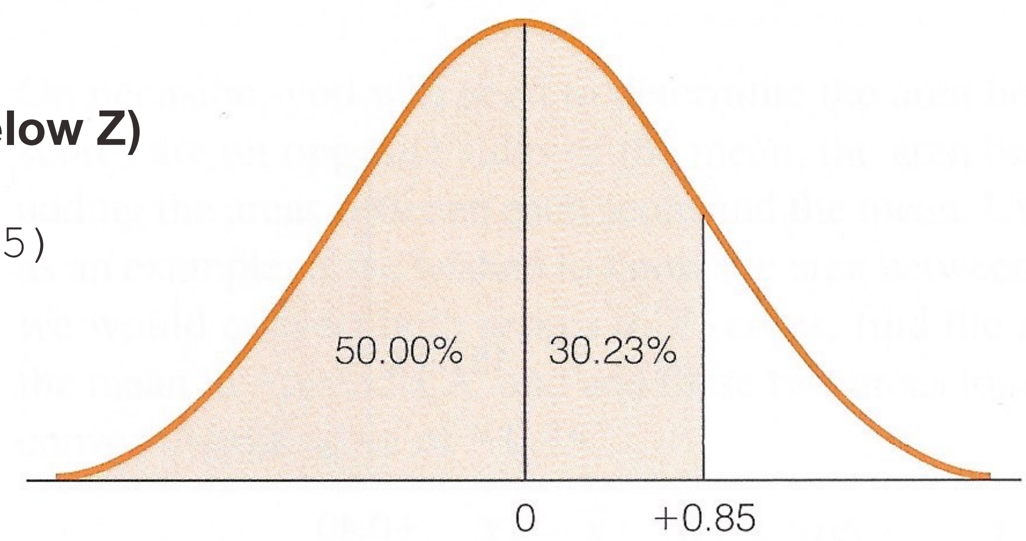
Source: Healey 2015, Appendix A, p.443.

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
```



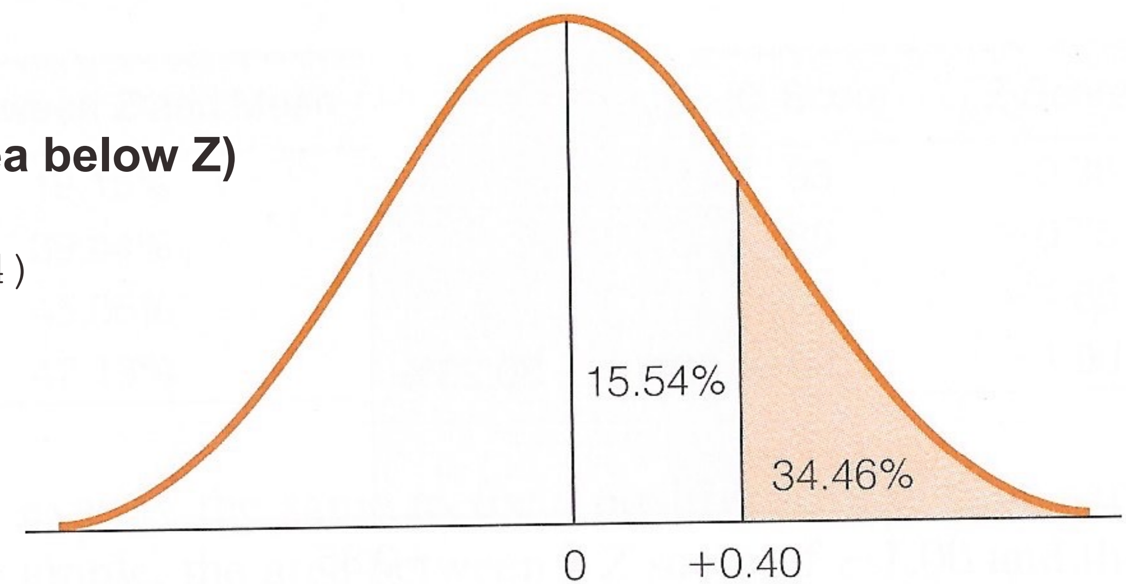
Source: Healey 2015, p.129.

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)

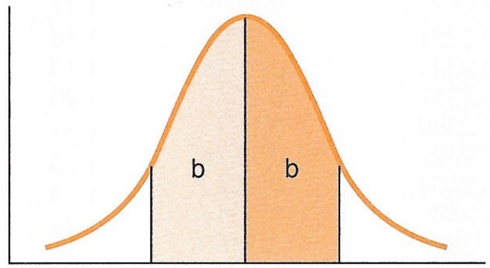
```
di 1-normal(0.4)

.34457826
```



Negative score

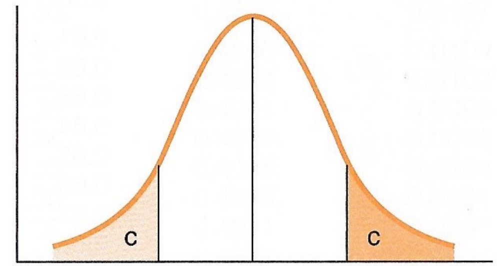
- Find your Z score in Column A
- To find area below a negative score
  - Look in column c
- To find area above a negative score
  - Add column b area to 0.50

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
...	...	...

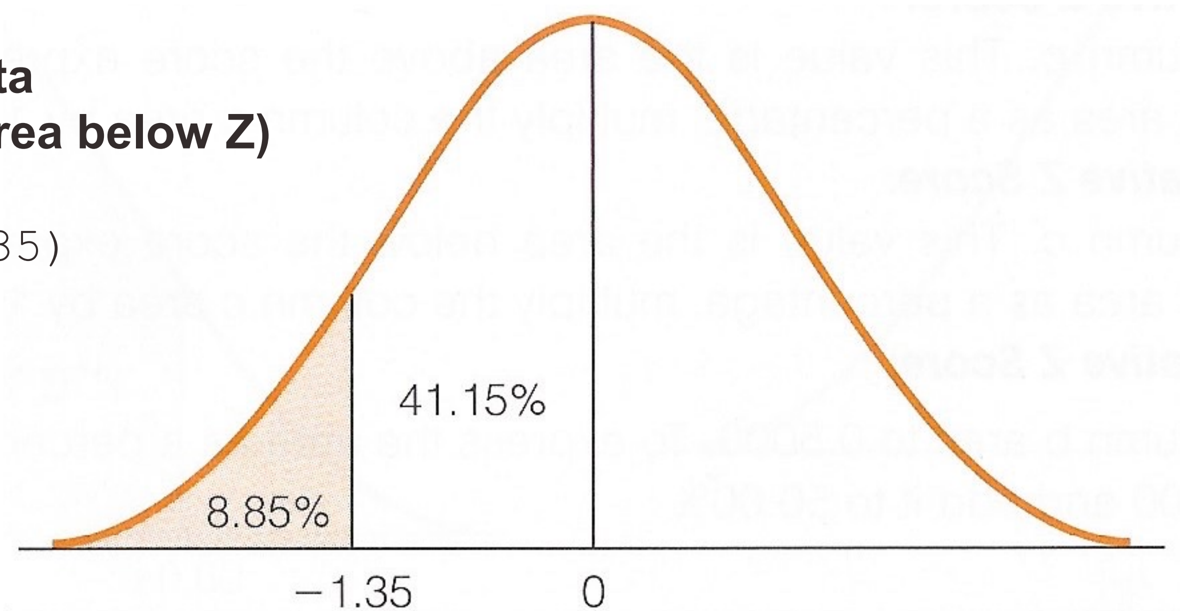
Source: Healey 2015, Appendix A, p.443.

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*

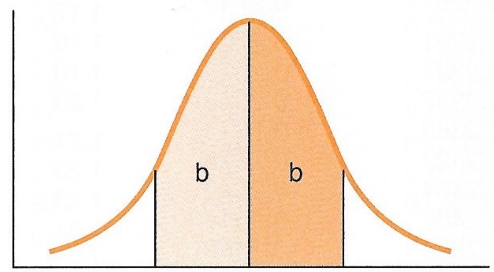
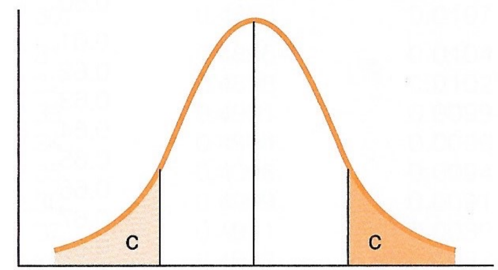


FIGURE A.2 Area Beyond *Z*



(a) *Z*	(b) Area Between Mean and *Z*	(c) Area Beyond *Z*	(a) *Z*	(b) Area Between Mean and *Z*	(c) Area Beyond *Z*
0.00	0.0000	0.5000	0.21	0.0832	0.4168
0.01	0.0040	0.4960	0.22	0.0871	0.4129
0.02	0.0080	0.4920	0.23	0.0910	0.4090
0.03	0.0120	0.4880	0.24	0.0948	0.4052
0.04	0.0160	0.4840	0.25	0.0987	0.4013
0.05	0.0199	0.4801	0.26	0.1026	0.3974
0.06	0.0239	0.4761	0.27	0.1064	0.3936
0.07	0.0279	0.4721	0.28	0.1103	0.3897
0.08	0.0319	0.4681	0.29	0.1141	0.3859
0.09	0.0359	0.4641	0.30	0.1179	0.3821
0.10	0.0398	0.4602	0.31	0.1217	0.3783
0.11	0.0438	0.4562	0.32	0.1255	0.3745
0.12	0.0478	0.4522	0.33	0.1293	0.3707
0.13	0.0517	0.4483	0.34	0.1331	0.3669
0.14	0.0557	0.4443	0.35	0.1368	0.3632
0.15	0.0596	0.4404	0.36	0.1406	0.3594
0.16	0.0636	0.4364	0.37	0.1443	0.3557
0.17	0.0675	0.4325	0.38	0.1480	0.3520
0.18	0.0714	0.4286	0.39	0.1517	0.3483
0.19	0.0753	0.4247	0.40	0.1554	0.3446
0.20	0.0793	0.4207	...	...	...

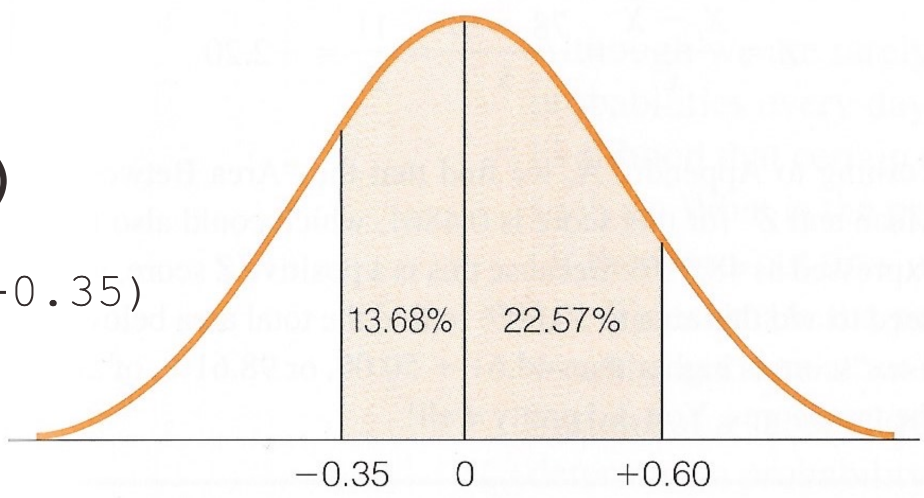
Source: Healey 2015, Appendix A, p.443.

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

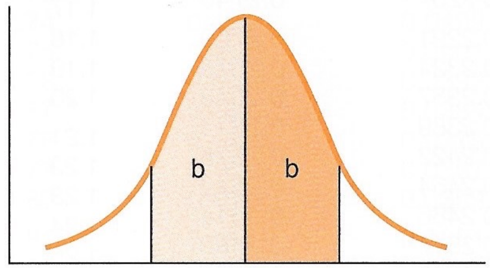
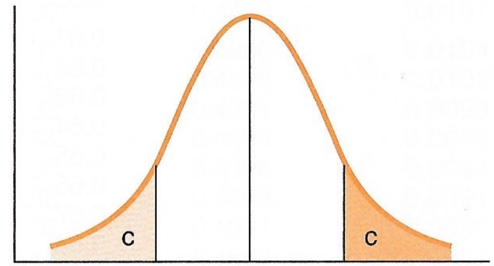


FIGURE A.2 Area Beyond Z



(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
...	...	...

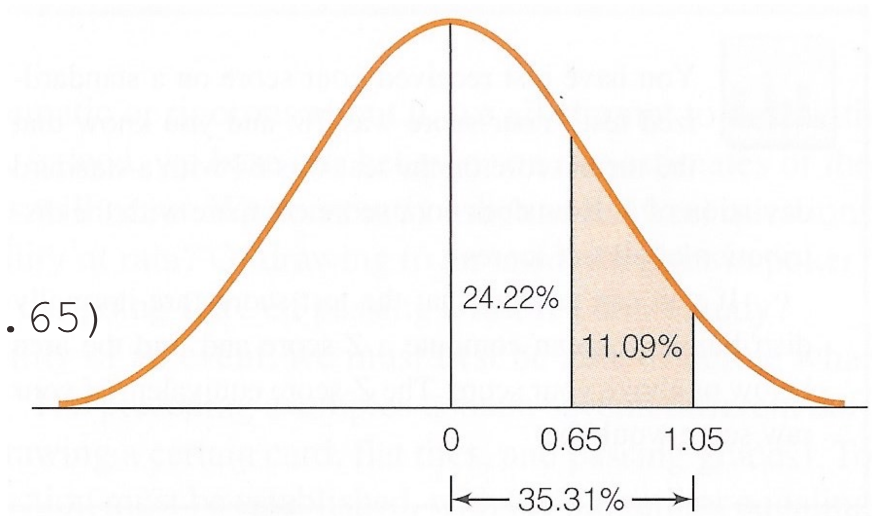
Source: Healey 2015, Appendix A, p.443.

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

Estimated date of delivery, 2017

Probability up to April 03

`z1=(277-281)/13`

`di normal(-0.31)`

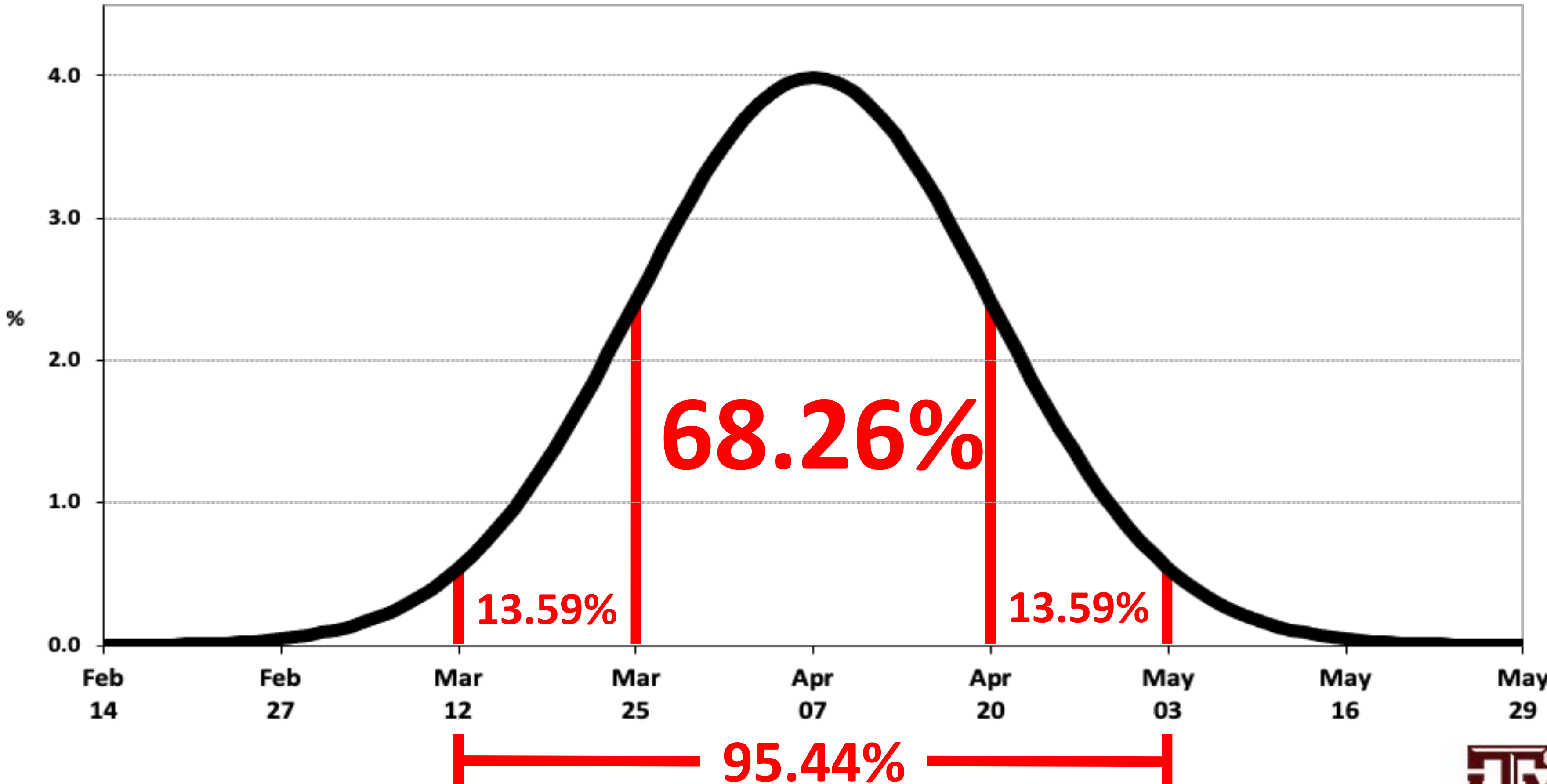
p = 0.3782805 = 37.83%

Probability between April 02–03

`z1=(277-281)/13; z2=(276-281)/13`

`di normal(-0.31)-normal(-0.38)`

p = 0.0263078 = 2.63%



Mean = 281 days; Std.Dev. = 13 days (based on Naegele's rule)

Estimated date of delivery, 2023

Probability up to June 30

`z1=(242-281)/13`

`di normal(-3)`

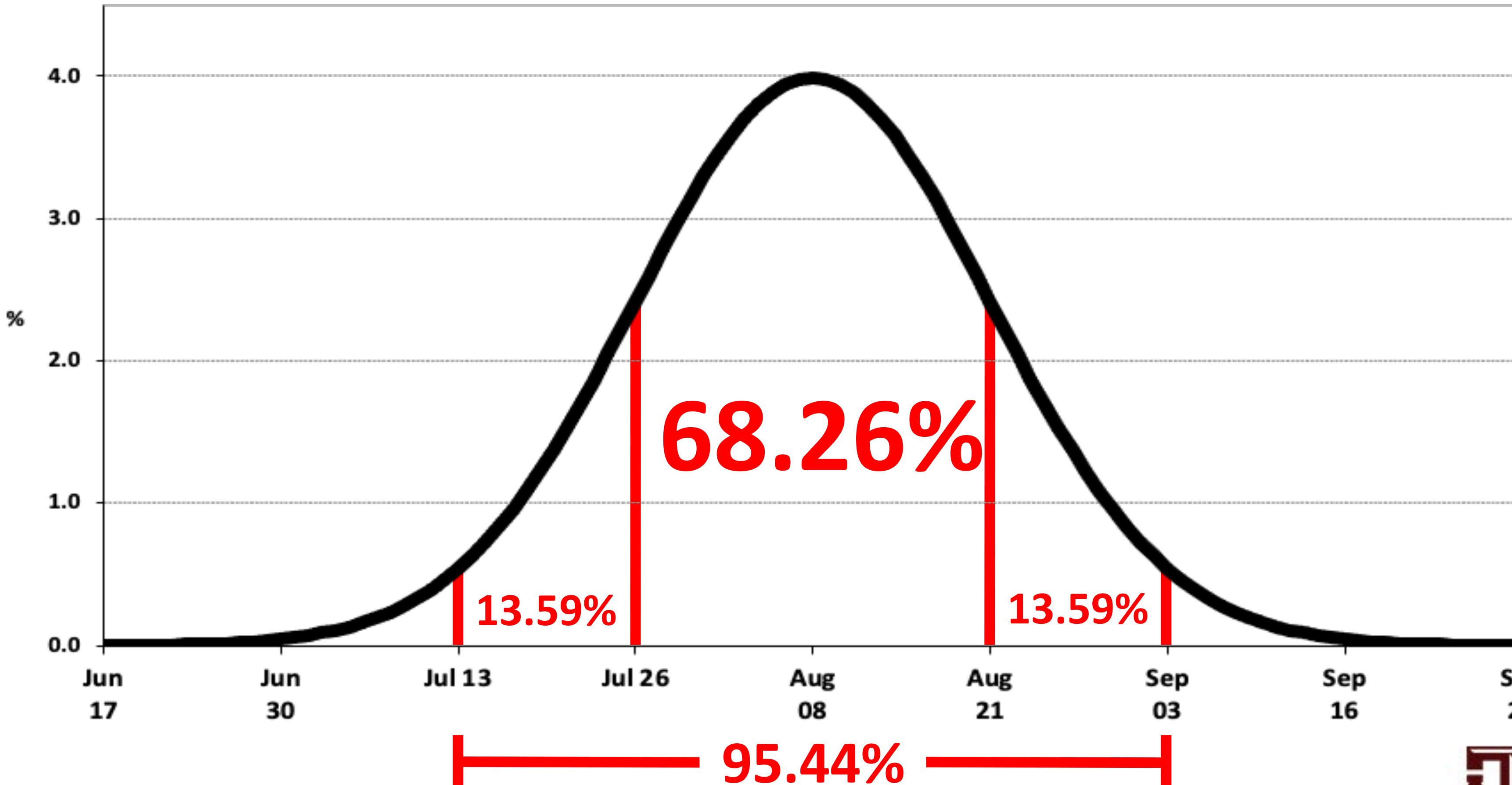
p = 0.0013499 = 0.14%

Probability between June 29–30

`z1=(242-281)/13; z2=(241-281)/13`

`di normal(-3)-normal(-3.08)`

p = 0.0003149 = 0.03%



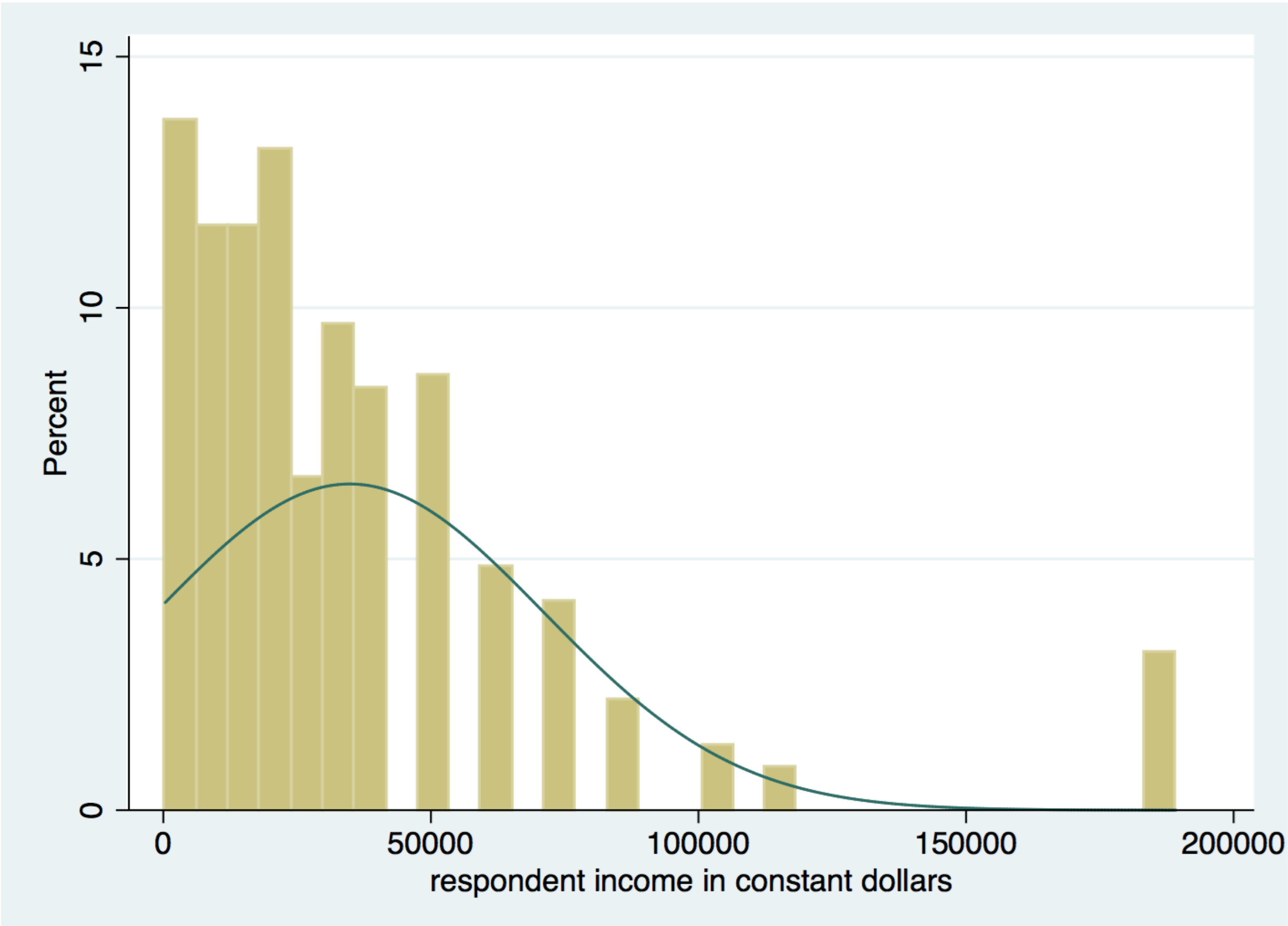
Mean = 281 days; Std.Dev. = 13 days (based on Naegele's rule)

Determining normality

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

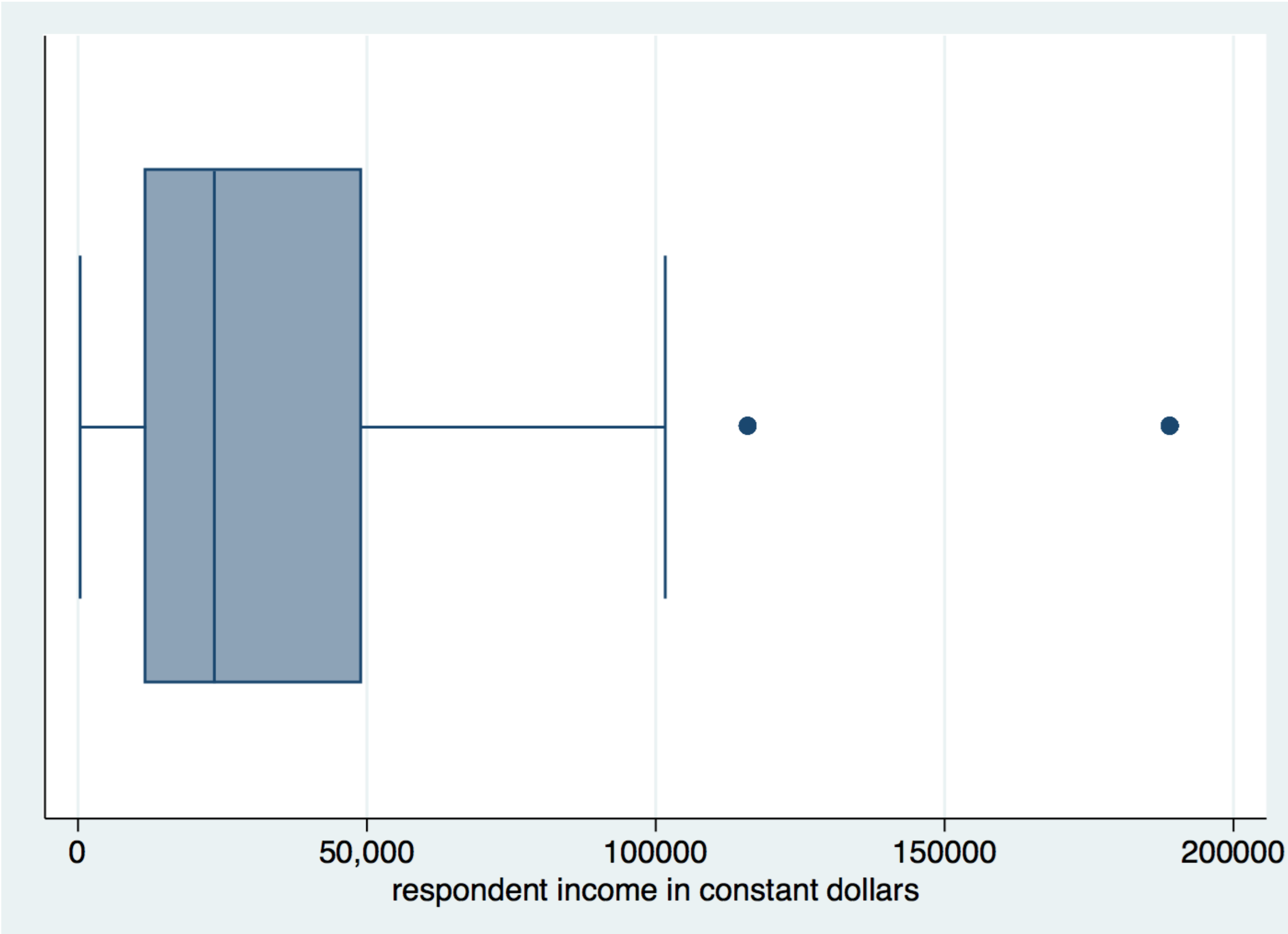
- We can analyze histograms, boxplots, outliers, quantile-normal plots to determine if variables have a normal distribution

Histogram of income



Source: 2016 General Social Survey.

Boxplot of income

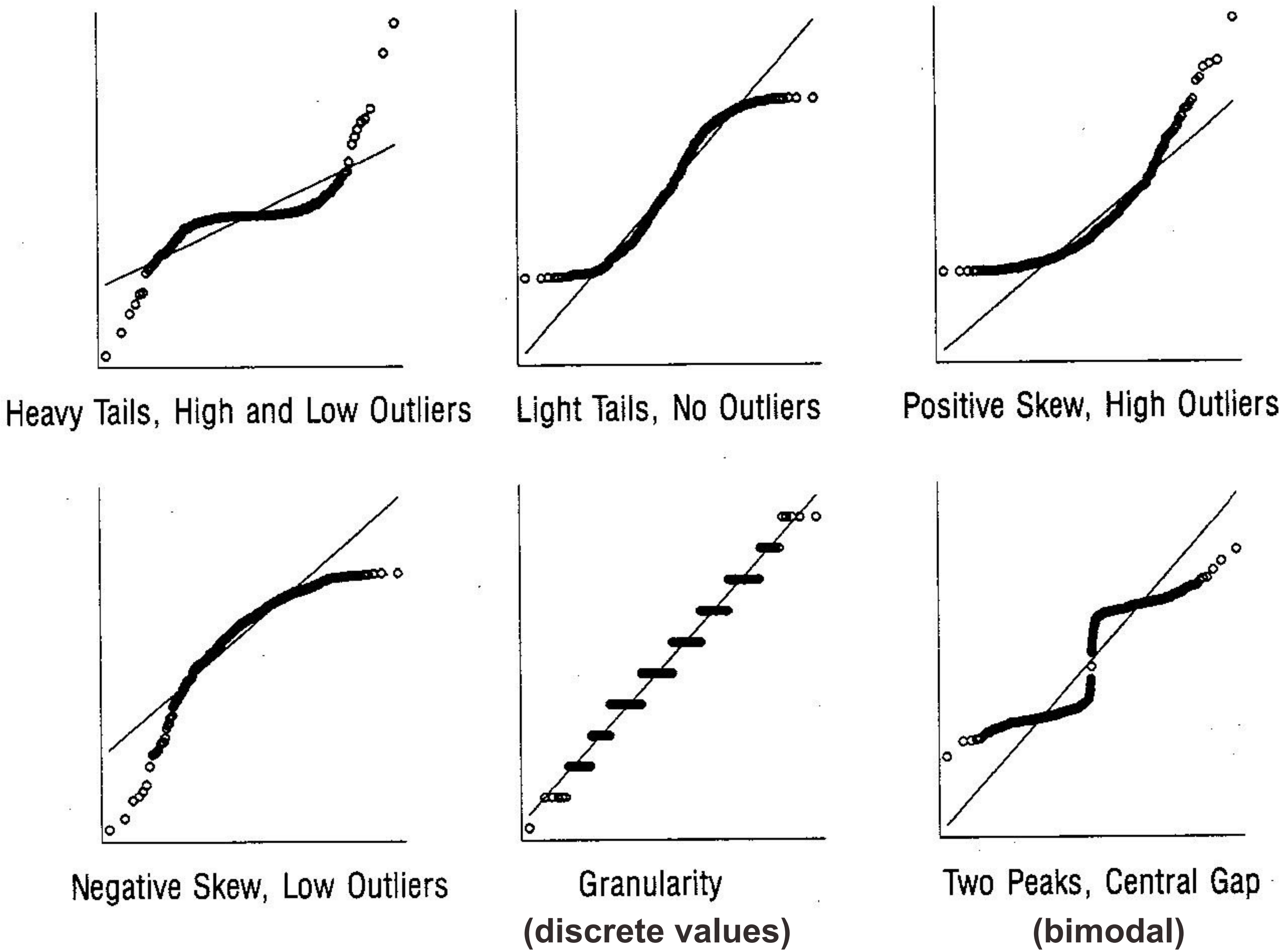


Source: 2016 General Social Survey.

Quantile-normal plots

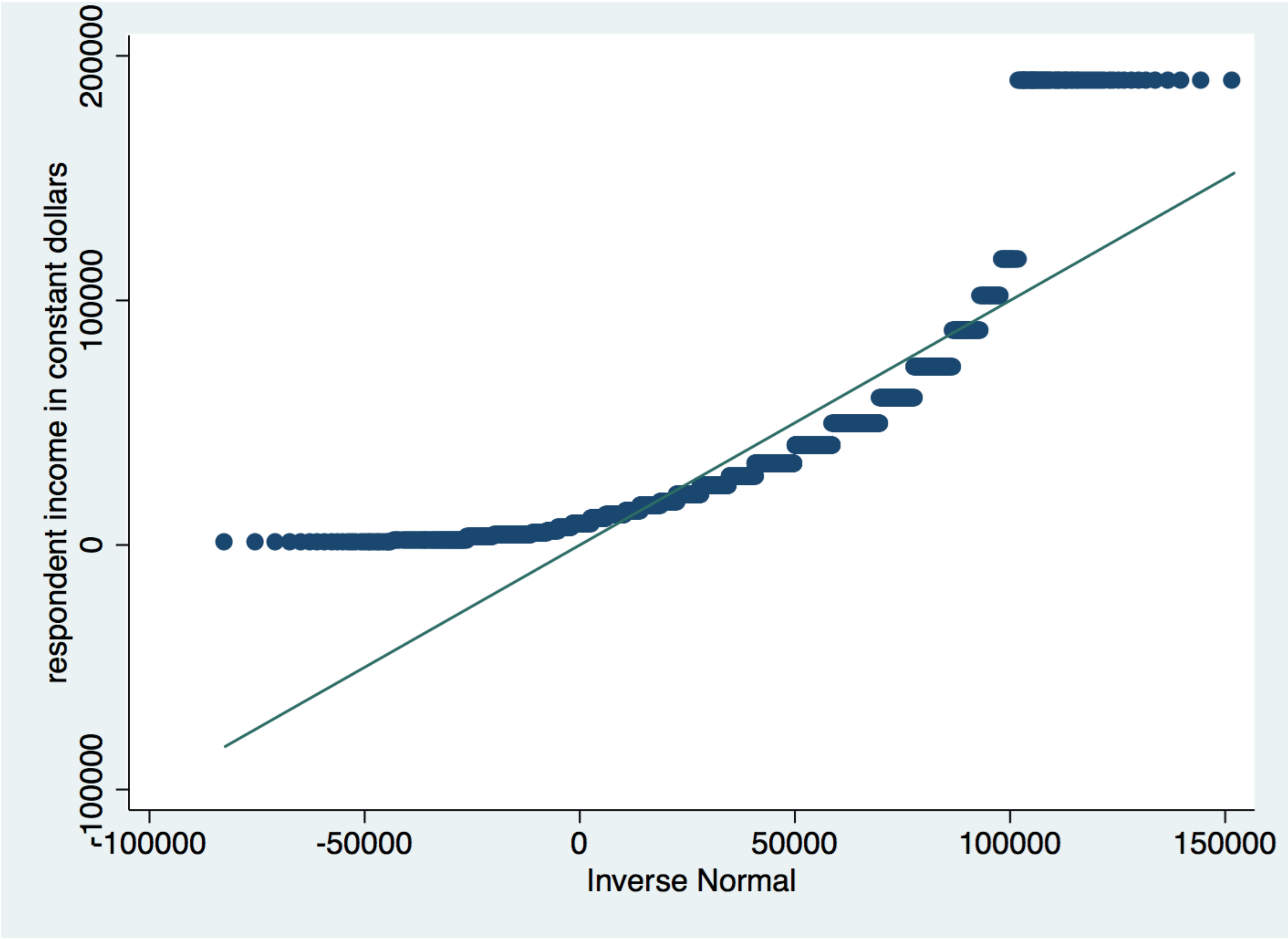
- A quantile-normal plot is a scatter plot
  - One axis has quantiles of the original data
  - The other axis has quantiles of the normal distribution
- If the points do not form a straight line or if the points have a non-linear symmetric pattern
  - The variable does not have a normal distribution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



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)

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

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

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

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

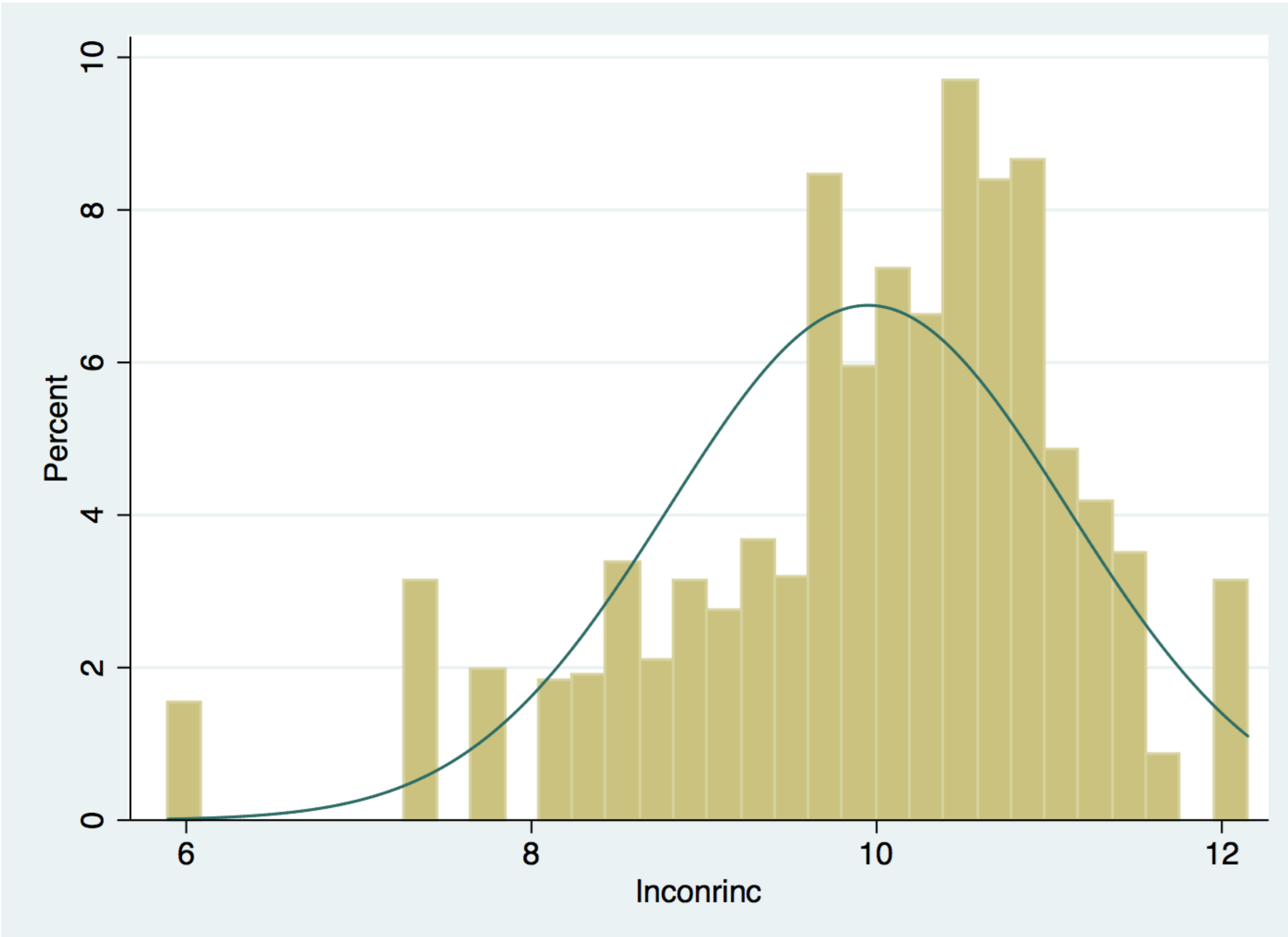
$$\log(Y) \longrightarrow q = 0$$

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

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

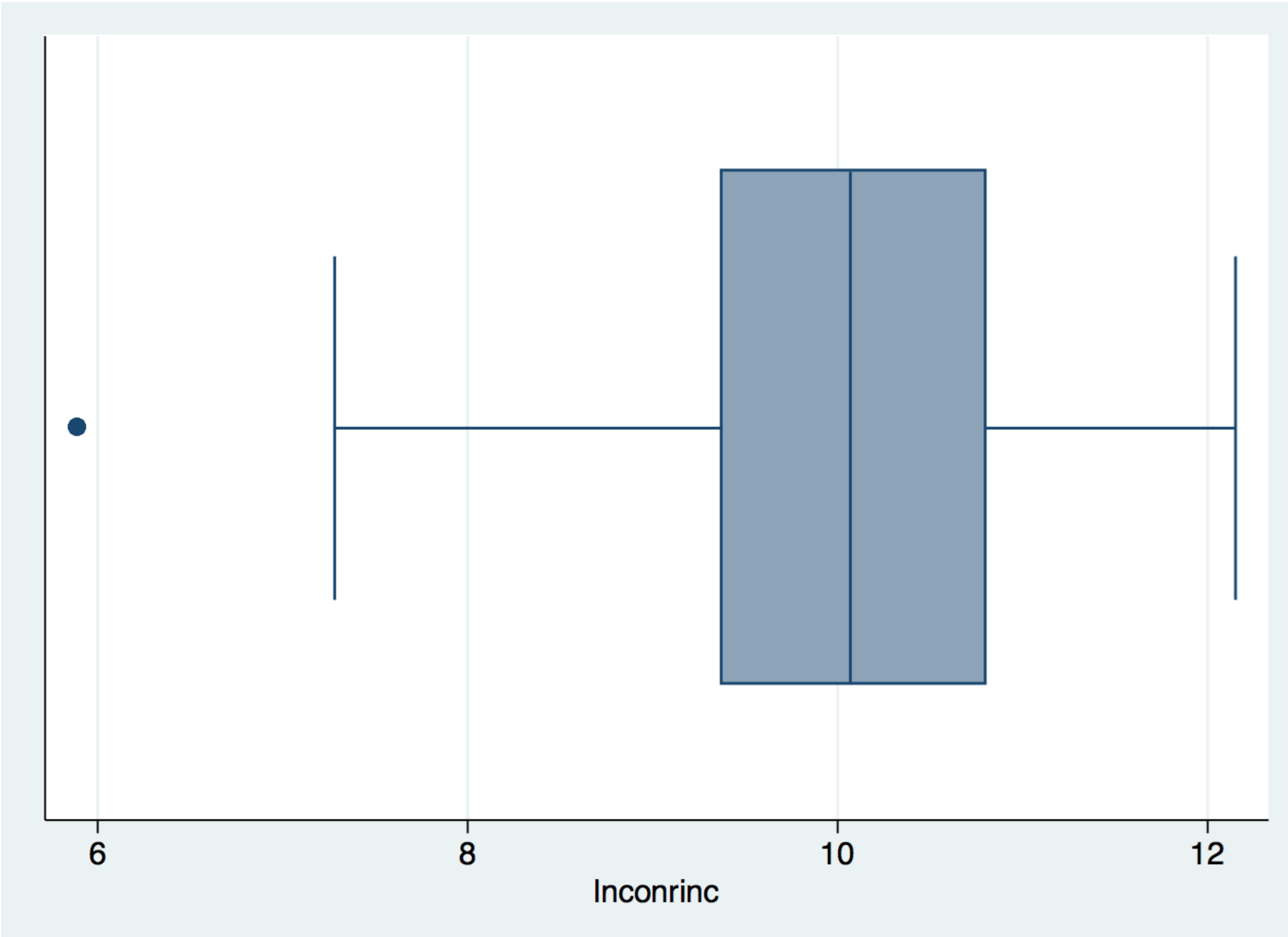
- $q>1$: reduce concentration on the right (reduce negative skew)
- $q=1$: original data
- $q<1$: reduce concentration on the left (reduce positive skew)
- $log(x+1)$ may be applied when $x=0$. If distribution of $log(x+1)$ is normal, it is called lognormal distribution

Histogram of log of income



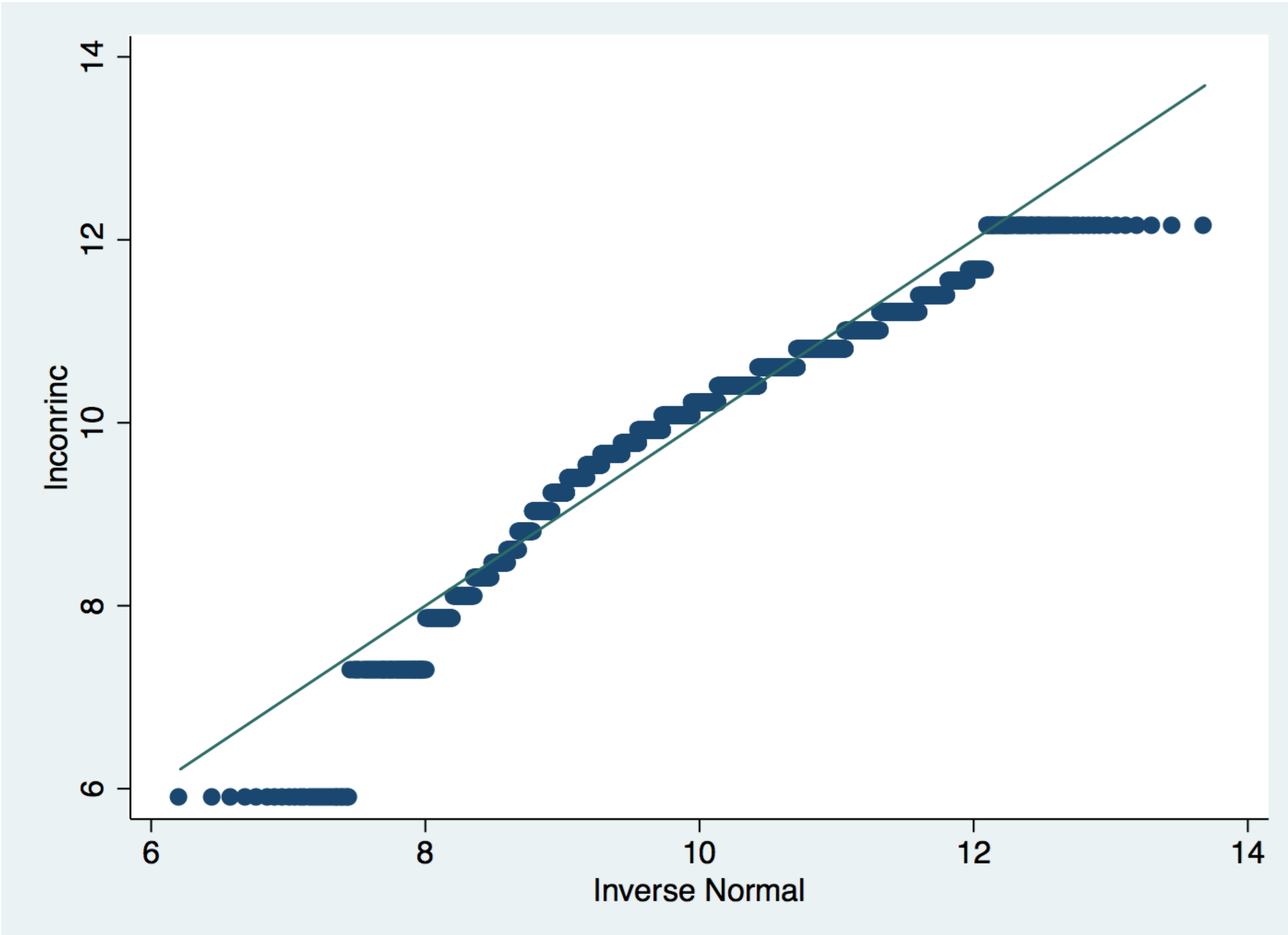
Source: 2016 General Social Survey.

Boxplot of log of income



Source: 2016 General Social Survey.

Quantile-normal plot of log of income

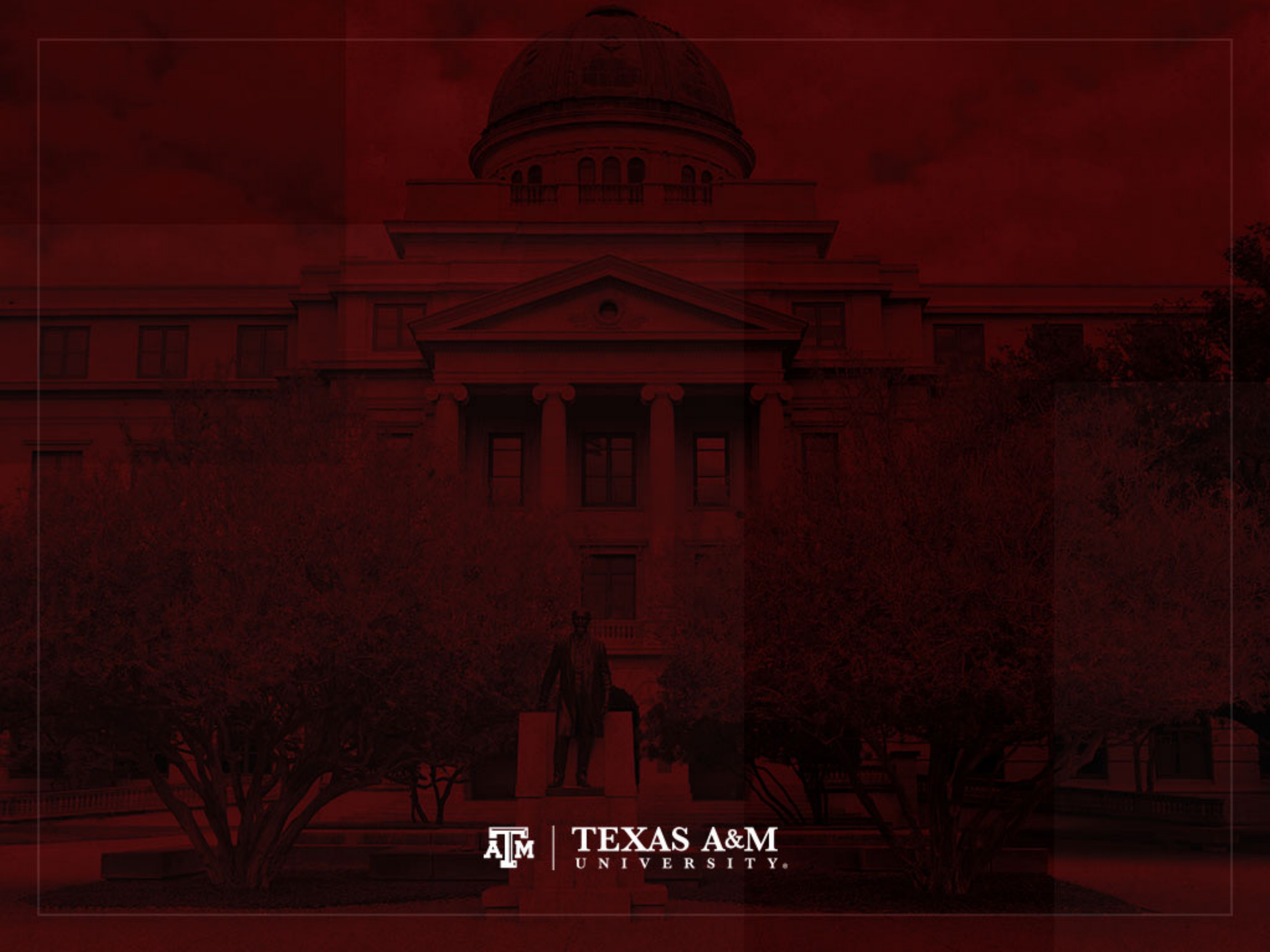


Source: 2016 General Social Survey.

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