Statistics Refresher Course

Day 1

Thomas Elliott

- 7th year grad student
- Interests: Social Movements, Sexualities, Quantitative Methodologies
- Email: telliott@uci.edu
- Office: SSPA 4166

Slides Available Online

- thomaselliott.me/bootcamp.html
- I'll try to have the next day's slides up the night before

What is Statistics?

- Statistics is used to summarize and analyze data
- Descriptive Statistics - summarizes data
- Inferential Statistics - makes predictions based on available data

Data

- Consists of some number of *variables* (k) about some number of *cases* (n)

Case

- Focus or subject of analysis
- Often people
- But can be anything
  - Organizations, countries, schools, friendship ties, etc.

Variables

- Characteristics about the cases relevant to our research question
- A person's age, education, income
- An organization's size, budget, location
- A nation's GDP, population, economic development

Data

- Often represented in a spreadsheet with cases as rows and variables as columns

id	age	sex	income	educ
1	45	male	57	16
2	32	female	36	14
3	27	male	17	12
4	56	male	89	18
5	37	female	104	16
6	41	female	95	24

Levels of Measurement

- Four levels:
  - Nominal - categorical
  - Ordinal - ordered categories
  - Interval - continuous with no zero
  - Ratio - continuous with zero

Nominal

- Categorical, not ordered
- Gender (male, female, etc)
- Religion (Protestant, Catholic, Jewish, Muslim, Atheist)
- State of Residence

Ordinal

- Categorical, Ordered
- Class (lower, working, middle, upper)
- Approval (strongly disapprove, disapprove, neither, approve, strongly approve)
- Rank (1st place, 2nd place, 3rd place)

Interval

- Interval between numbers is meaningful
- But no meaningful zero
- Temperature
- Hours on a 12 hr clock

Ratio

- Continuous variable
- Zero is meaningful - means there is nothing
- Age
- Years of education
- Income

Descriptive Statistics

- Summarize Data
- Mean, Median, Mode
- Graphs and Figures
- Standard Deviation and Variance

Measures of Central Tendency

- Mean, Median, and Mode
- Describes the center of the distribution of one variable for a number of cases
- What does a typical case look like?

Mean

- AKA the average
- The sum of the observations divided by the total number of observations

$$\bar{y} = \frac{y_1 + y_2 + y_3 + \ldots + y_n}{n}$$

$$\bar{y} = \frac{\sum y}{n}$$

y
3
0
4
9
5
8
2
6
3
9

n=10

$$\bar{y} = \frac{3+0+4+9+5+8+2+6+3+9}{10}$$

$$\bar{y} = 4.9$$

Practice!

- You are a TA and the instructor has asked you to calculate the mean score for your students, which are below:

98	86	90	72	85	87	92	78	95	89	90	81

98	86	90	72	85	87	92	78	95	89	90	81

n=12

$$\bar{y} = \frac{98 + 86 + 90 + 72 + 85 + 87 + 92 + 78 + 95 + 89 + 90 + 81}{12}$$

$$\bar{y} = 86.92$$

Mean

- Drawback
  - Susceptible to unusually high or unusually low numbers

Mean

- Say you are at a meeting with 30 people, who all make around $50K
- Chancellor Drake walks in ($400K)
- Now the average salary is $61K

Median

- Middle number, when observations are sorted
- For odd number of cases, just the middle number
- For even number of cases, the average of the two middle numbers

y	sorted
3	0
0	2
4	3
9	3
5	4
8	5
2	6
6	8
3	9
9	9

$$\frac{4+5}{2} = 4.5$$

Practice!

- Now the instructor wants to know the median score for your students

98	86	90	72	85	87	92	78	95	89	90	81

98	86	90	72	85	87	92	78	95	89	90	81

72	78	81	85	86	87	89	90	90	92	95	98

$$\frac{87+89}{2} = 88$$

Mean vs Median

- Mean Household Income: $32,195
- Median Household Income: $26,672

Mode

- The observation that appears the most often

y
3
0
4
9
5
8
2
6
3
9

Mode: 3, 9

Practice!

- What is the mode of the test scores?

98	86	90	72	85	87	92	78	95	89	90	81

98	86	90	72	85	87	92	78	95	89	90	81

98	86	90	72	85	87	92	78	95	89	90	81

Mode: 90

Graphs and Tables

- Graphs and tables are common tools used to describe data
- Can represent data visually

Data

- General Social Survey, 1972-2010
- Respondent's Religion

Frequency Table

	Frequency	Percentage
Protestant	32,556	59.34
Catholic	13,482	24.57
Jewish	1,127	2.05
None	5,726	10.44
Other	1,970	3.59
Total	54,861	100

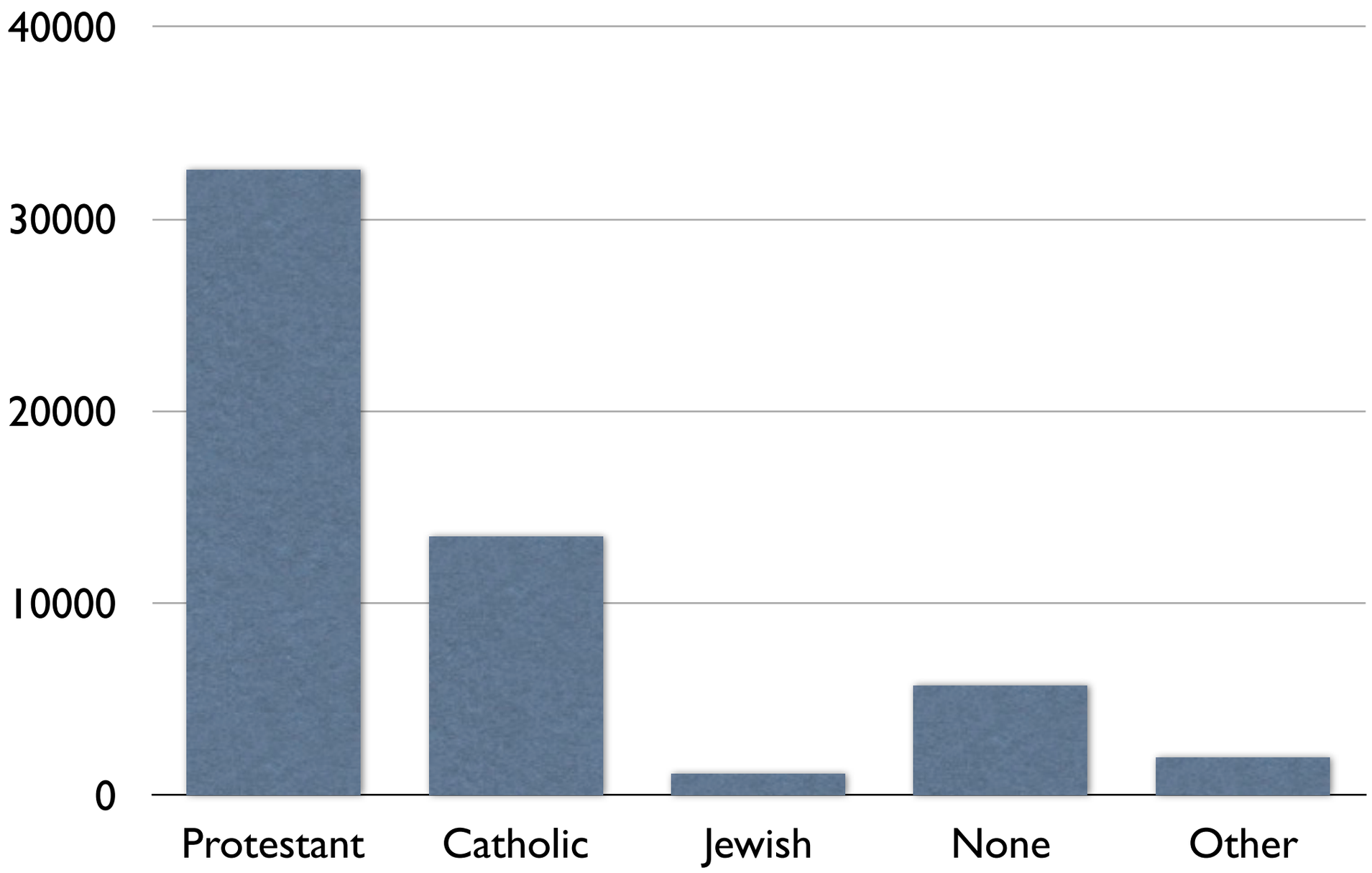
Frequency Table

Frequency is the number of people who answered "Protestant"

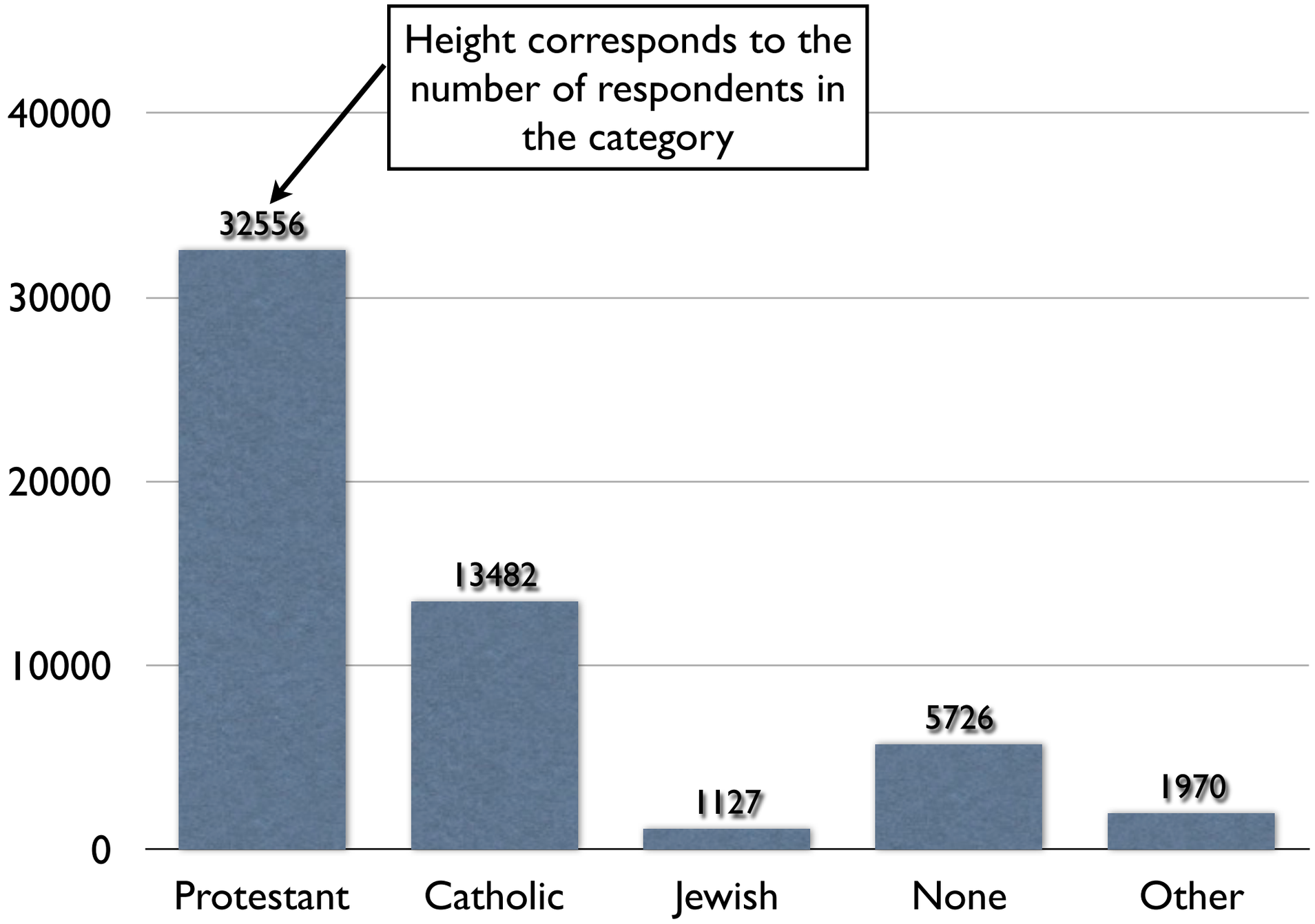
Percentage is the frequency divided by the total

	Frequency	Percent
Protestant	32,556	59.34
Catholic	13,482	24.57
Jewish	1,127	2.05
None	5,726	10.44
Other	1,970	3.59
Total	54,861	100

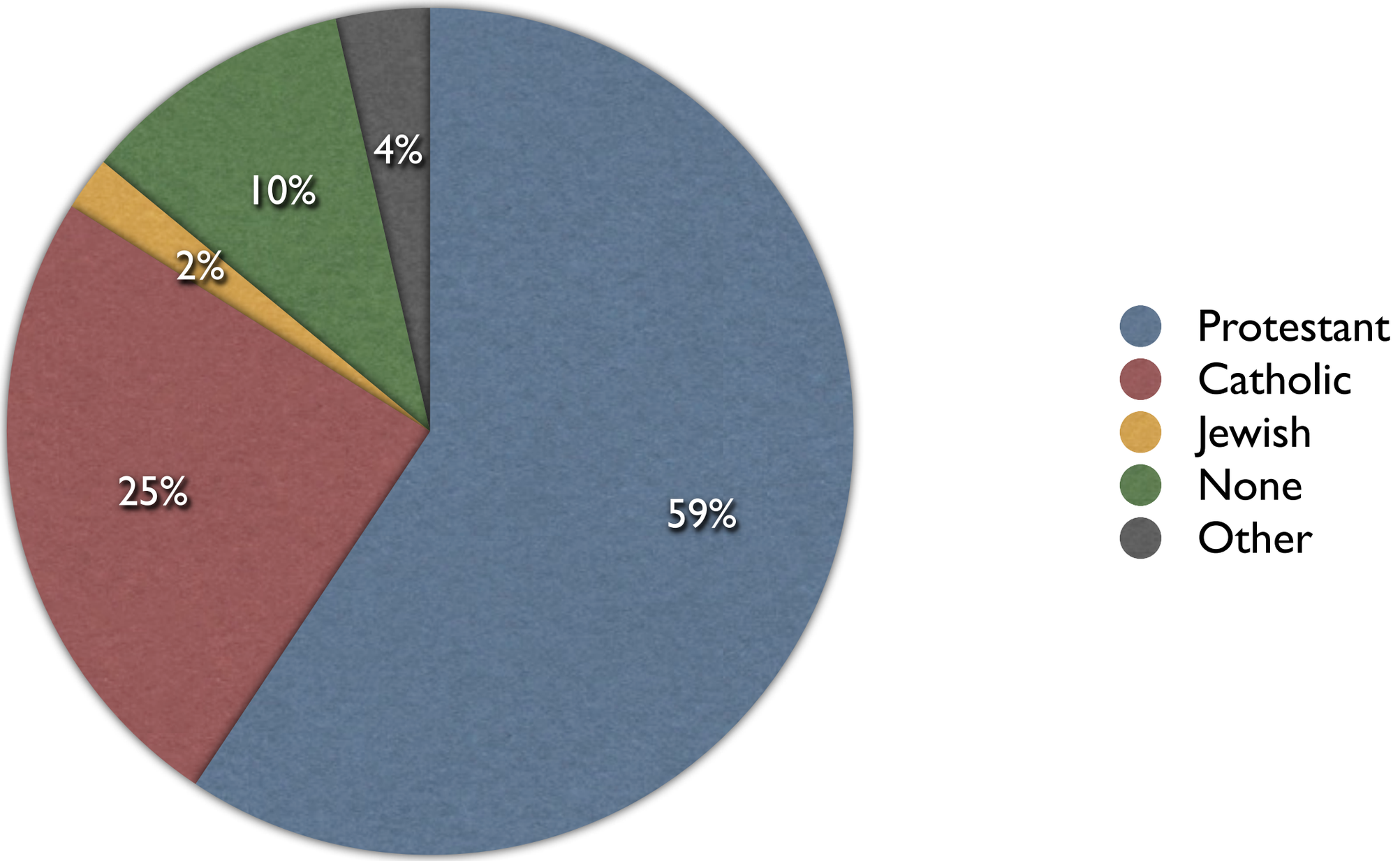
Bar Chart


40000
30000
20000
10000
0
Protestant
Catholic
Jewish
None
Other

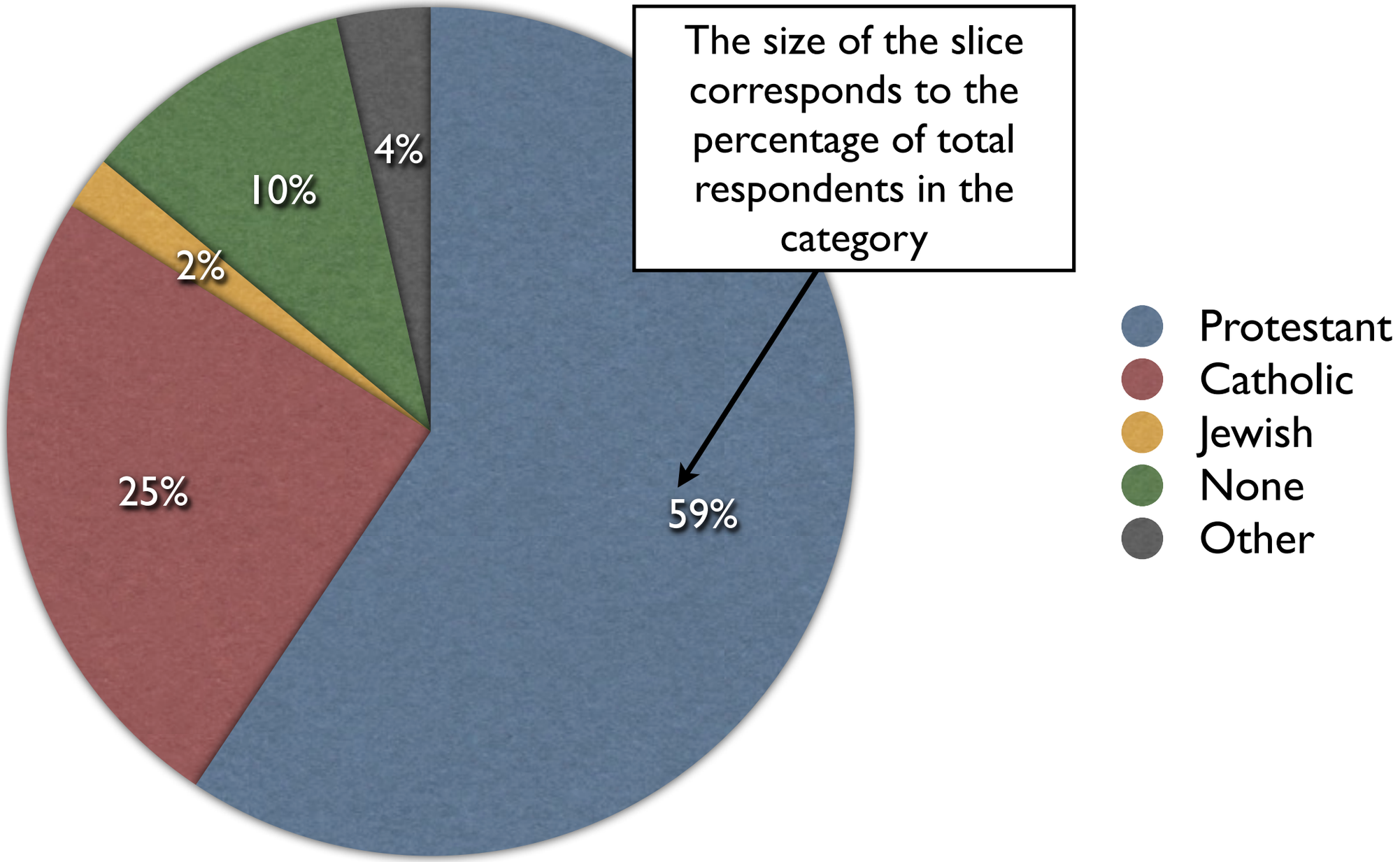
Bar Chart


Height corresponds to the number of respondents in the category
40000
30000
20000
10000
0
32556
13482
1127
5726
1970
Protestant
Catholic
Jewish
None
Other

Pie Chart



Pie Chart


The size of the slice corresponds to the percentage of total respondents in the category
59%
25%
2%
10%
4%
Protestant
Catholic
Jewish
None
Other

Central Tendency

- Mean and Median tell us about the center of the distribution

5	4	4	3	5	3

$$\bar{y} = \frac{5+4+4+3+5+3}{6} = 4$$

1	12	2	1	10	-2

$$\bar{y} = \frac{1+12+2+1+10-2}{6} = 4$$

Central Tendency

- So mean doesn't tell us about how spread out the observations are

Variance

- Tells us how spread out the data are
- Approximately the average of the squared deviations

$$s^2 = \frac{\sum (y_i - \bar{y})^2}{n - 1}$$

Deviation

- How far away a single observation is from the average

$$y_i - \bar{y}$$

12	4	7	5	6	2

$$\bar{y} = \frac{12 + 4 + 7 + 5 + 6 + 2}{6} = 6$$

6	-2	1	-1	0	-4

What happens if we summed the deviations?

$$6 - 2 + 1 - 1 + 0 - 4 = 0$$

Summed Deviations

- Deviations from the mean always sum to zero
- So if we want a measure of the spread, we need to square them before we sum them

12	4	7	5	6	2

6	-2	1	-1	0	-4

36	4	1	1	0	16

$$36 + 4 + 1 + 1 + 0 + 16 = 58$$

Variance

- Now we divide by the number of observations minus one (for reasons we'll talk about later)

$$\frac{58}{6-1} = 11.6$$

Variance

$$s^2 = \frac{\sum (y_i - \bar{y})^2}{n - 1}$$

Variance

- What are the units of variance?
- Original units squared
  - So if we are measuring age in years, variance is in years^2
- Solution: take the square root of variance

Standard Deviation

- The positive square root of variance

$$s = \sqrt{s^2} = \sqrt{\frac{\sum (y_i - \bar{y})^2}{n-1}}$$

Standard Deviation

- The positive square root of variance

$$s = \sqrt{s^2} = \sqrt{\frac{\sum (y_i - \bar{y})^2}{n-1}}$$

$$s = \sqrt{11.6} = 3.41$$

Practice!

- The instructor for the class you are TAing for now wants the standard deviation of the test scores

98	86	90	72	85	87	92	78	95	89	90	81

98	86	90	72	85	87	92	78	95	89	90	81

$$\bar{y} = 86.9$$

11.1	-0.9	3.1	-14.9	-1.9	0.1	5.1	-8.9	8.1	2.1	3.1	-5.9

123.21	0.81	9.61	222.01	3.61	0.01	26.01	79.21	65.61	4.41	9.61	34.81

$$123.21 + 0.81 + 9.61 + 222.01 + 3.61 + 0.01 + 26.01 + 79.21 + 65.61 + 4.41 + 9.61 + 34.81 = 578.92$$

$$s^2 = \frac{578.92}{12 - 1} = 52.63$$

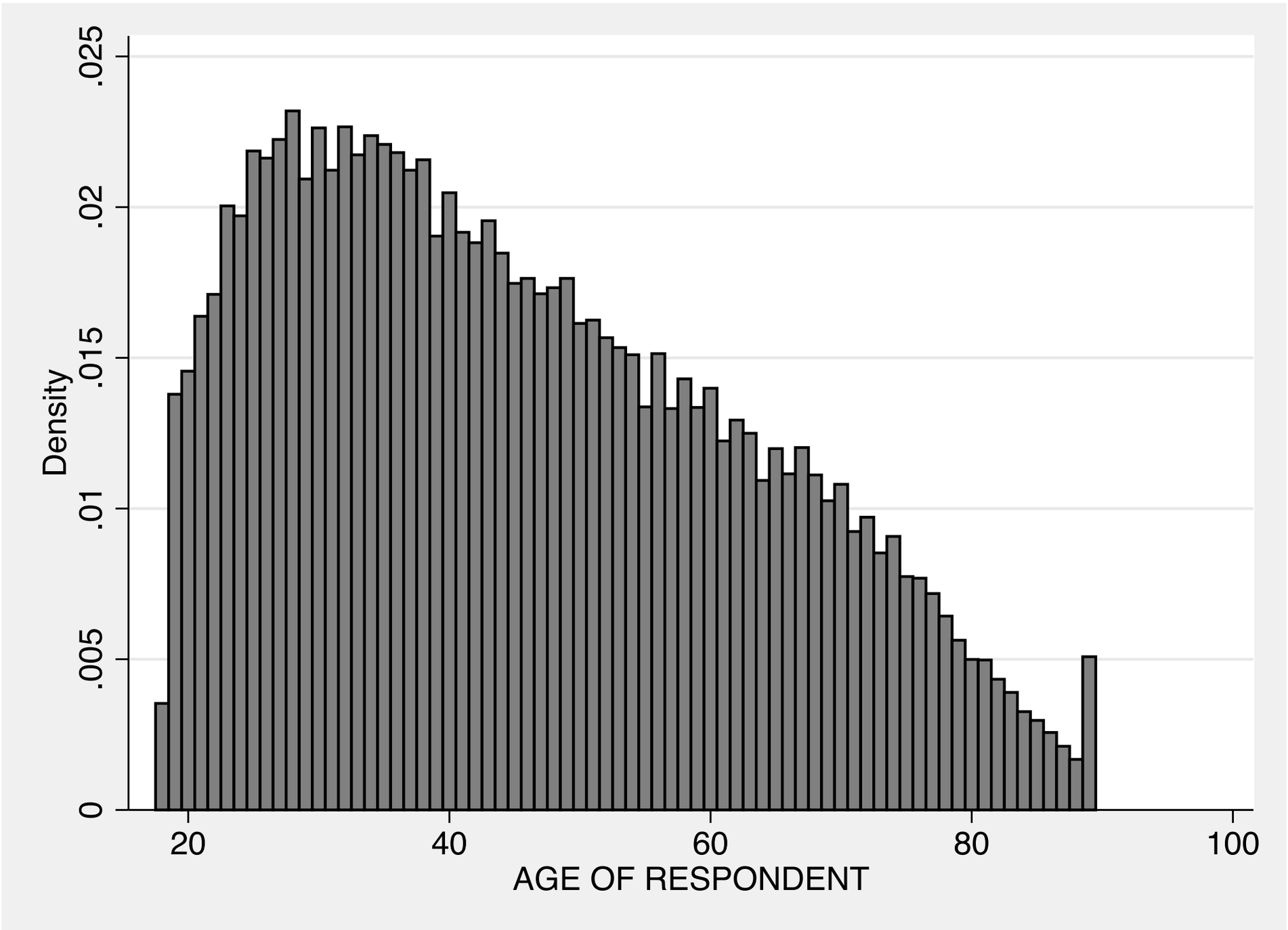
$$s = \sqrt{s^2} = \sqrt{52.63} = 7.25$$

Distributions

Distribution

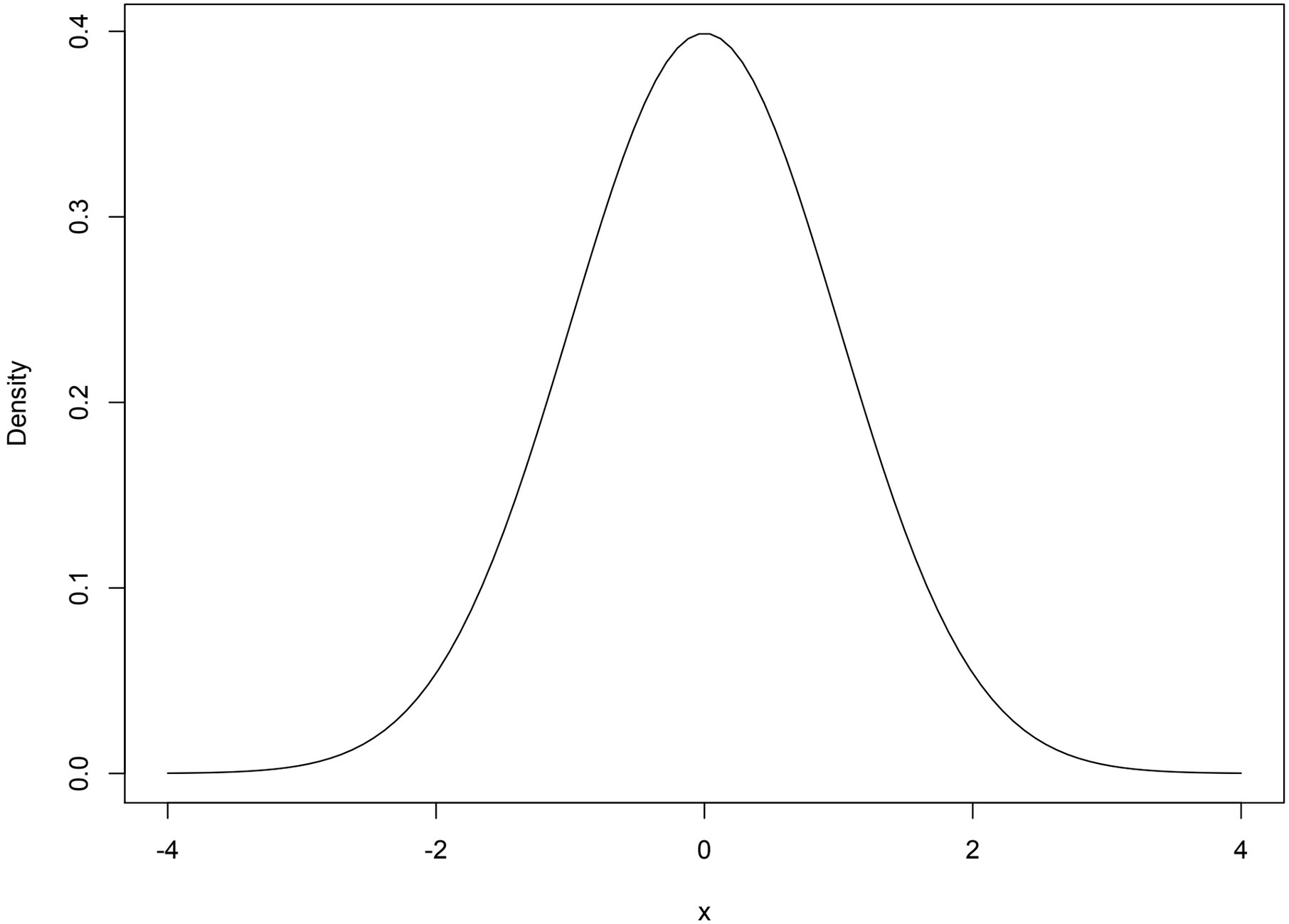
- A distribution describes how data are distributed across potential values
- We can graph this using histograms

Histogram



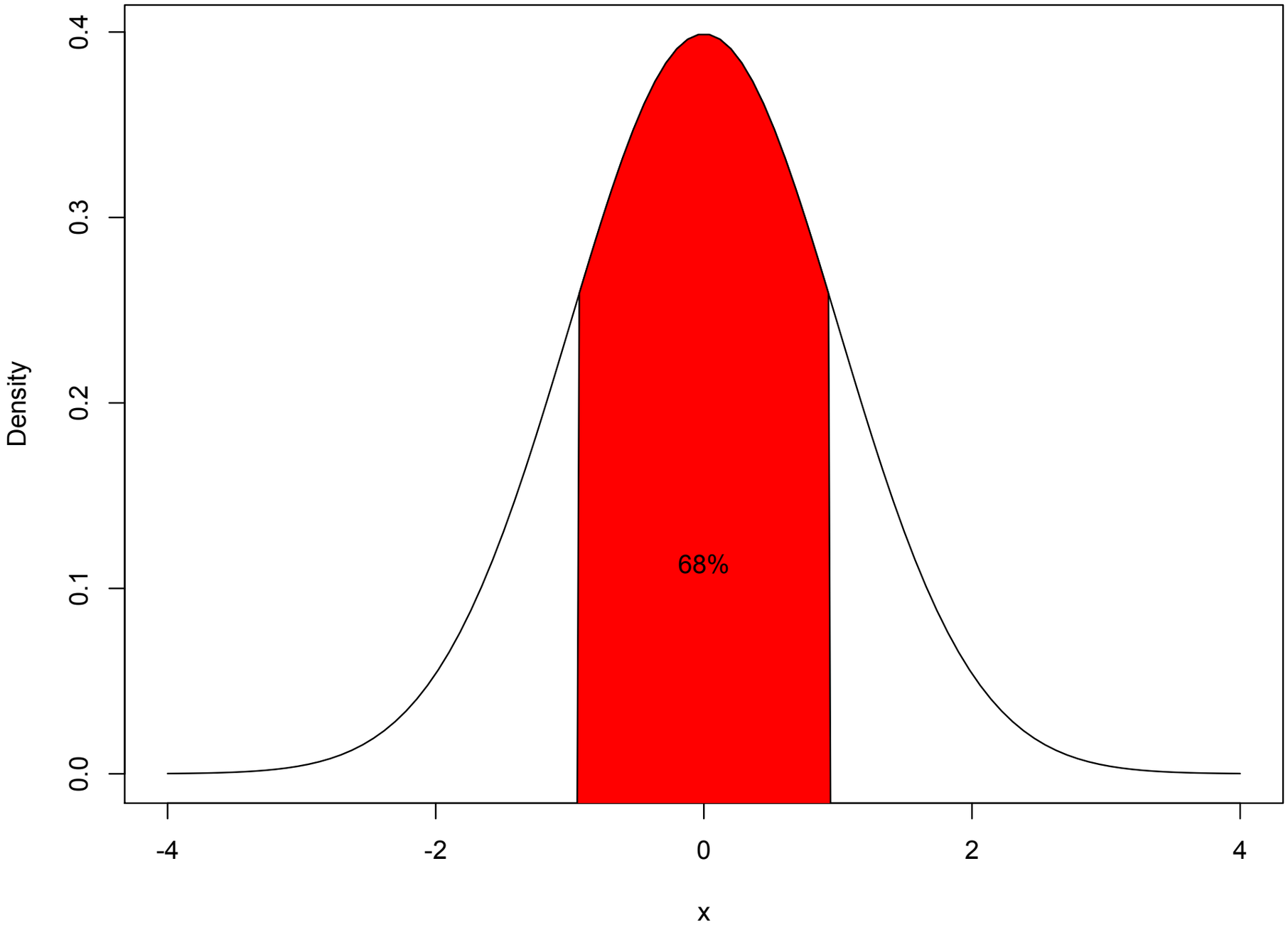
Normal Distribution

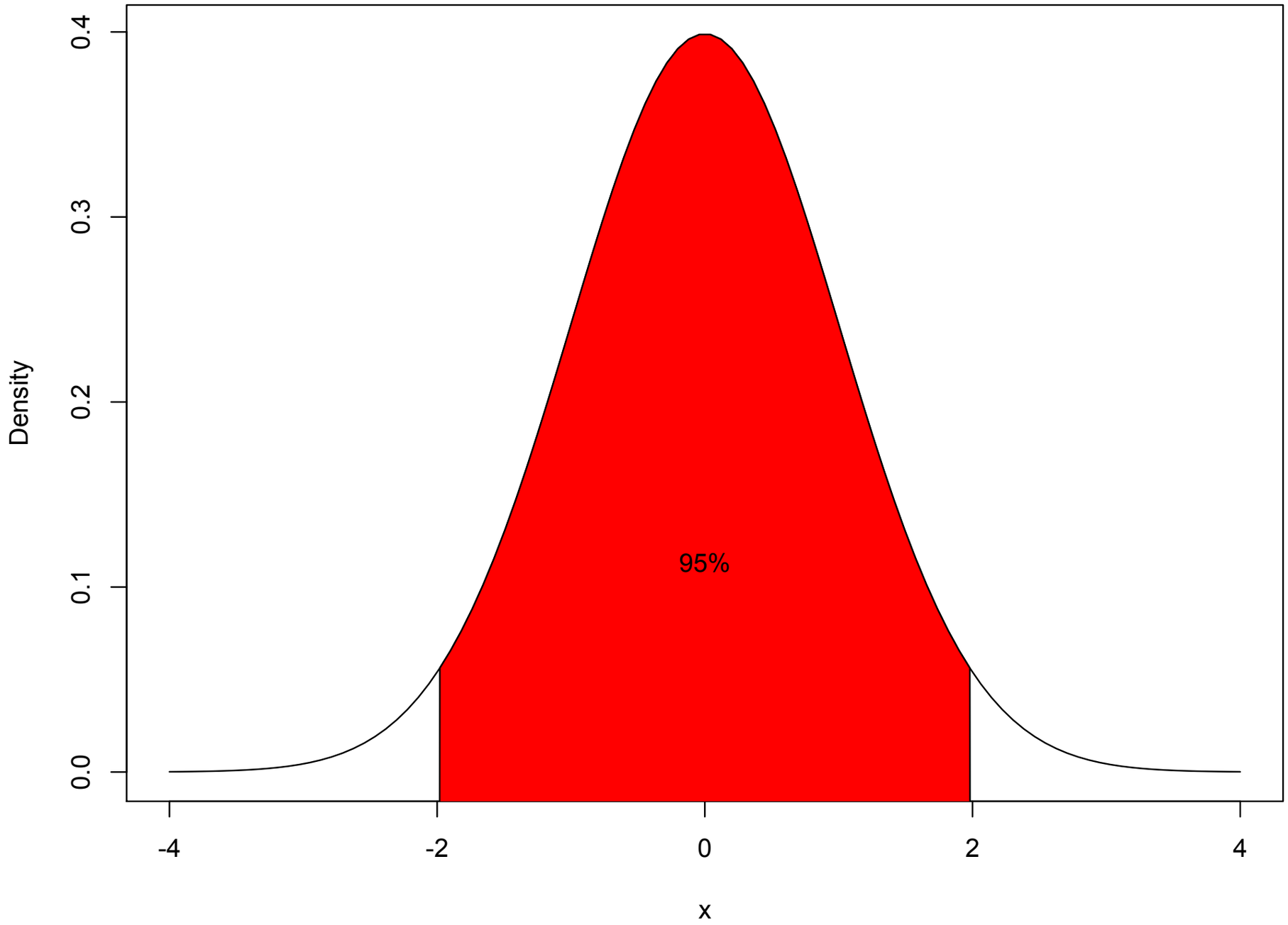
- The most common distribution in social science is the normal distribution
- Also called bell-shaped

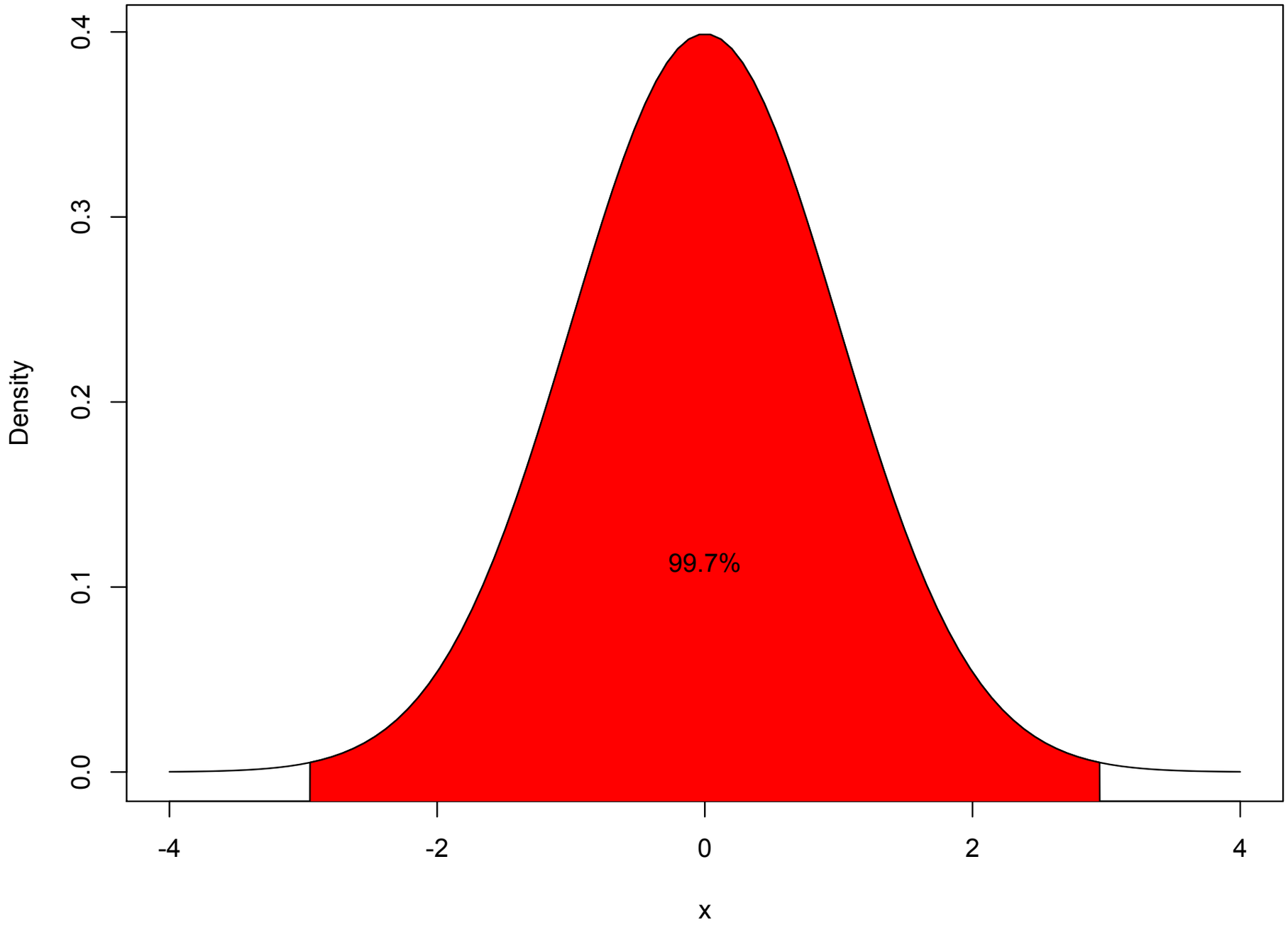

Density
0.0
0.1
0.2
0.3
0.4
-4
-2
0
2
4
x

Normal Distribution

- Properties:
  - Mean and median are the same
  - 68.2% of observations are within 1 standard deviation of the mean
  - 95.4% of observations are within 2 standard deviations of the mean
  - 99.7% of observations are within 3 standard deviations of the mean


Density
0.0
0.1
0.2
0.3
0.4
-4
-2
0
2
4
x
68%


Density
0.0
0.1
0.2
0.3
0.4
-4
-2
0
2
4
x
95%


Density
0.0
0.1
0.2
0.3
0.4
99.7%
-4
-2
0
2
4
x

Inferential Statistics

- Inferential Statistics involves making inferences about a population based on a sample of that population

Populations and Samples

- Population - everything that qualifies as a potential case
  - If you are studying US college students, population is ALL US college students
- Sample - those cases you collect data on
  - Use samples to make inferences about the population

Research Design

- When designing new research projects, defining your population, and constructing a valid sampling design, is crucial

Population Statistics

- We use different symbols to distinguish population statistics from sample statistics
- Mean:
  - Sample: $\bar{x}$
  - Population: μ
- Standard Deviation
  - Sample: s
  - Population: σ

Standard Deviation

- The equation of sample and population means are the same, but the equations for standard deviation are slightly different

Sample

$$s = \sqrt{\frac{\sum (x - \bar{x})^2}{n - 1}}$$

Population

$$\sigma = \sqrt{\frac{\sum (x - \mu)^2}{N}}$$

Population Statistics

- In general, we call population statistics **parameters**.
- Sample statistics are simple called **statistics**
- Parameters have one true value, but we rarely know what it is
- Statistics will have different values depending on the sample you take

Sample Distribution

- Let's say you are interested in a research question in which the population is all undergrads at UCI ~22,000 students
- Instead of interviewing all 22,000, you interview a sample of 100 students and ask about GPA

Sample Distribution

- According to UCI's website, the population mean GPA in Fall 2011 was 3.03
- You ask 100 students one day and your sample mean GPA was 2.96
- You ask 100 students the next day and your sample mean GPA was 3.05

Sample Distribution

- The sample mean and standard deviation of your sample will depend on who you interviewed
- Samples have natural variability, and the sample statistics will reflect that variability.

Sample Distribution

- There are lots and lots of possible samples of 100 students from 22,000 undergrads
- 1.5×10^{276} possible samples
- The sample statistics from all these samples produce a distribution of sample statistics
- This is the sample distribution

Mean of Sample Distribution

$$\mu_{\bar{X}} = \frac{\sum \bar{X}}{N}$$

$\mu_{\bar{X}}$ mean of the sampling distribution

$\bar{X}$ the sample mean

N the number of samples

Standard Error

- The standard deviation of the sampling distribution

$$\sigma_{\bar{X}} = \sqrt{\frac{\sum(X - \mu_{\bar{X}})^2}{N}}$$

Central Limit Theorem

- If you take multiple samples of size n from a population and compute their mean, the distribution of the computed means will be normally distributed for large numbers of samples

$$\mu_{\bar{X}} = \mu \qquad \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

Law of Large Numbers

- if n (the sample size) is large, the distribution of sample means will be approximately normally distributed

$$\mu_{\bar{X}} = \mu \qquad \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

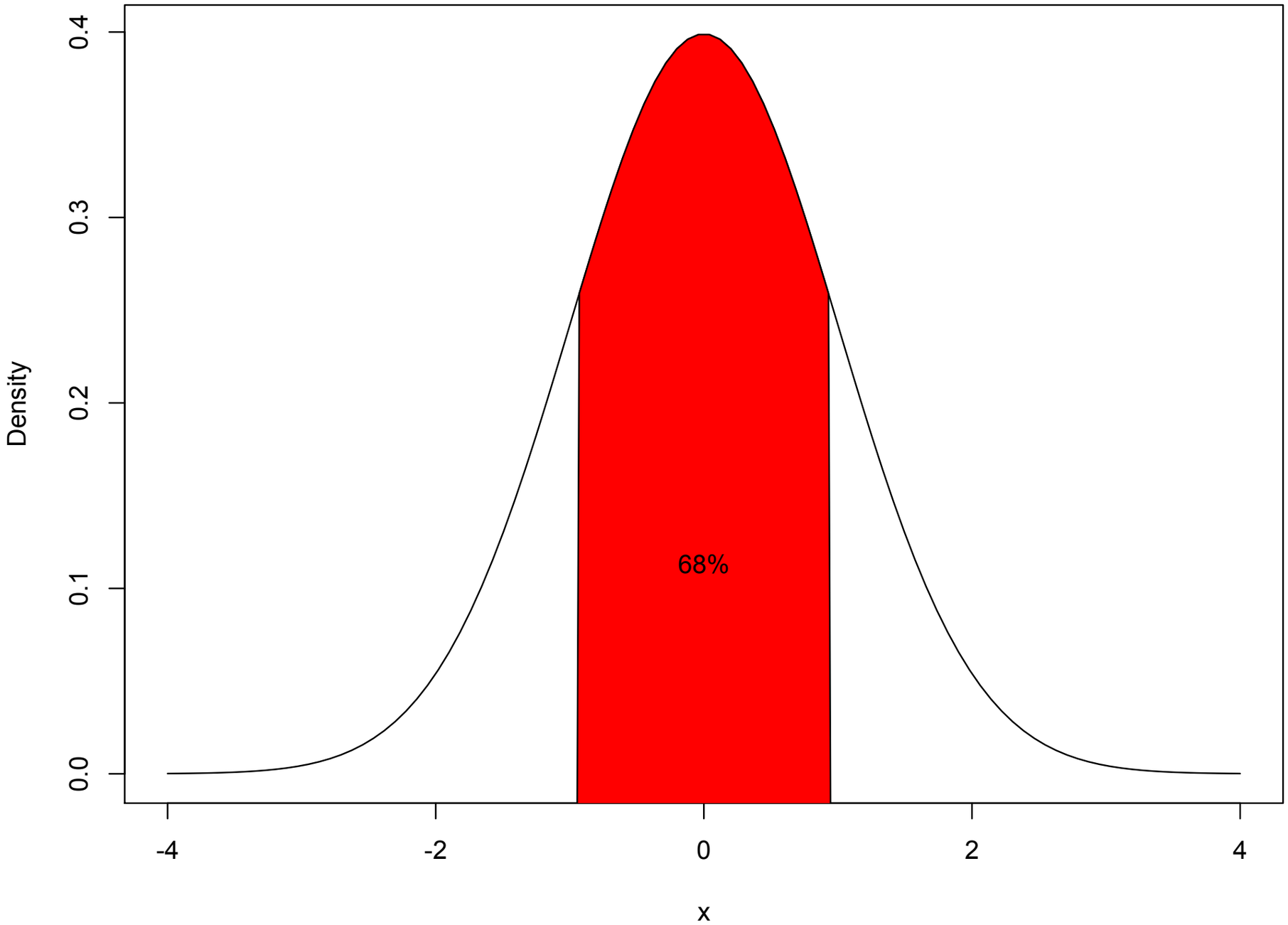
Sample Distribution

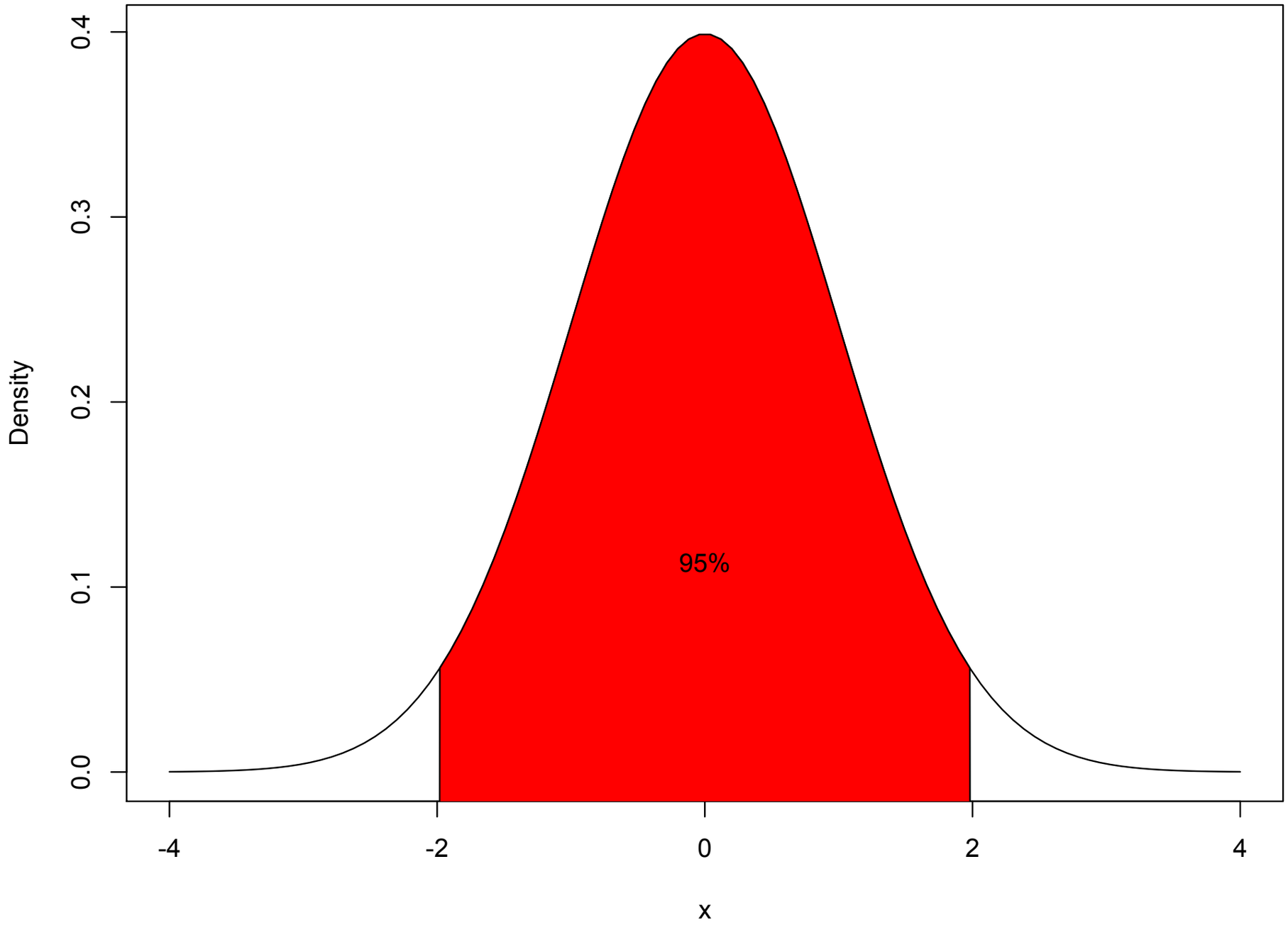
- These two theorems tell us that sampling distributions are normal

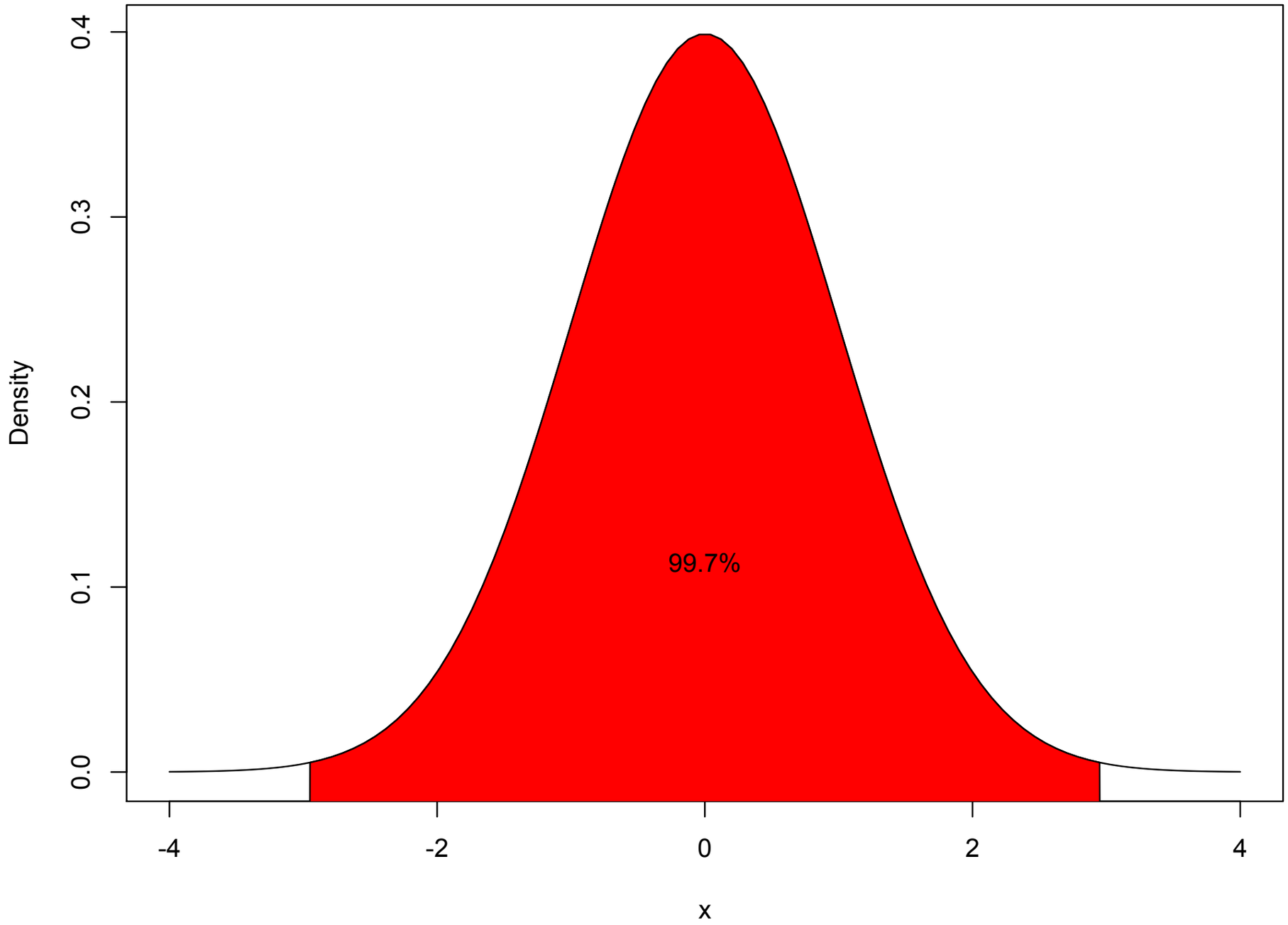
Normal Distribution

- Also known as the bell curve
- Mean and median are the same
- Set proportion of observations within standard deviations of the mean

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


Density
0.0
0.1
0.2
0.3
0.4
-4
-2
0
2
4
x
68%


Density
0.0
0.1
0.2
0.3
0.4
-4
-2
0
2
4
x
95%


Density
0.0
0.1
0.2
0.3
0.4
99.7%
-4
-2
0
2
4
x

Z scores

- Standardized values along a normal curve
- Essentially the ratio of an observation's deviance score to the average deviance score (standard deviation)

Z scores

Population

$$Z = \frac{x - \mu}{\sigma}$$

Sample

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

Z scores

- Given a set of observations with mean 4.6 and standard deviation 2.1, what is the Z score for an observation with value 3?

$$\bar{X} = 4.6 \qquad s = 2.1 \qquad X_i = 3$$

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

$$Z = \frac{3 - 4.6}{2.1}$$

$$Z = \frac{-1.6}{2.1}$$

$$Z = -0.76$$

Practice!

- Given a normally distributed set of observations with mean of 45 and standard deviation of 10, what is the z score of a value of 62?

$$\bar{X} = 45 \qquad s = 10 \qquad X_i = 62$$

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

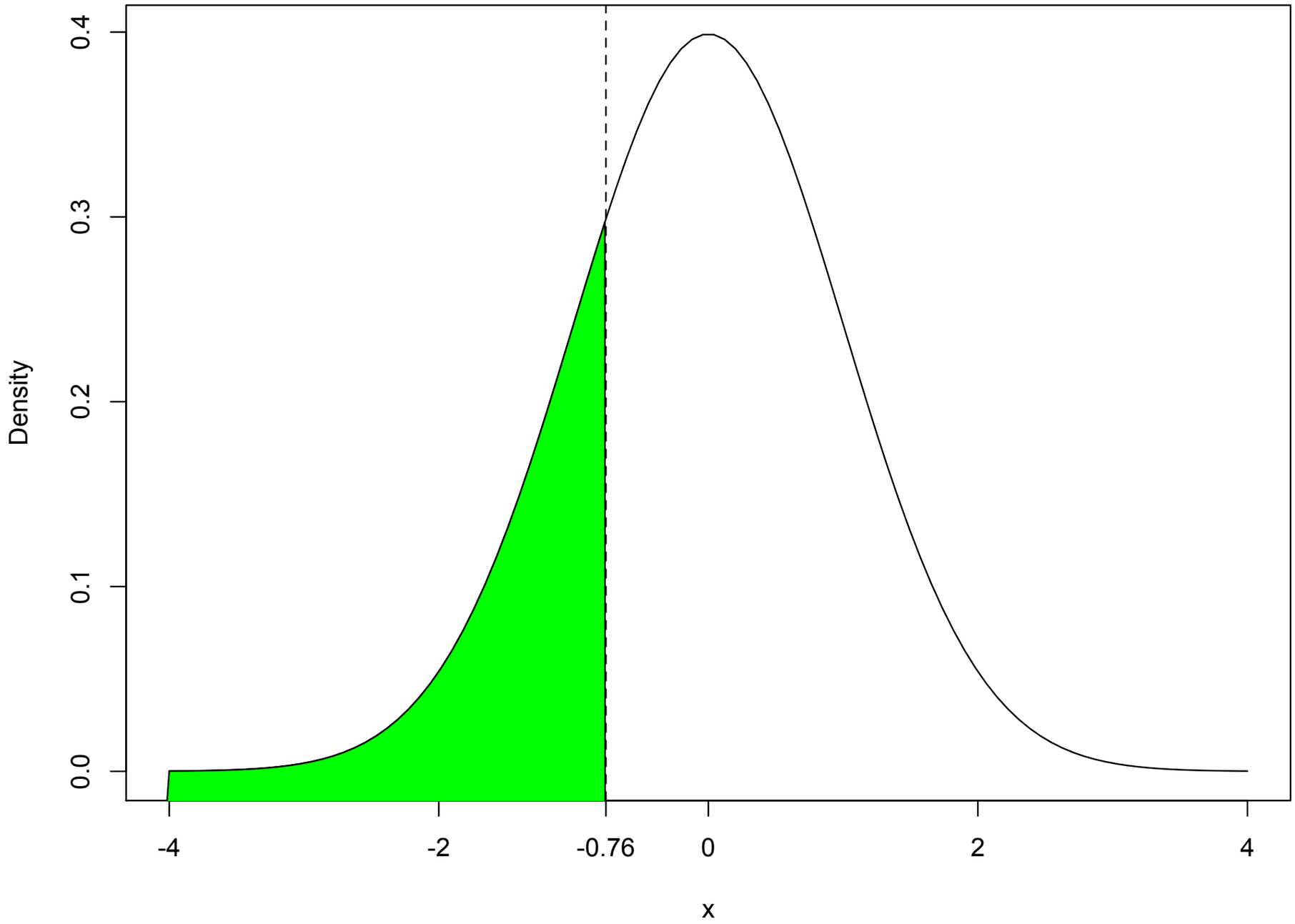
$$Z = \frac{62 - 45}{10}$$

$$Z = \frac{17}{10}$$

$$Z = 1.7$$

Z scores

- Z scores are normally distributed with mean of 0 and standard deviation of 1
- We can use this to figure out the proportion of observations that fall above or below a certain point

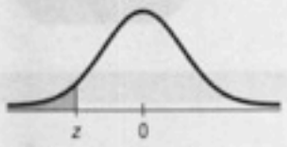

Density
0.0
0.1
0.2
0.3
0.4
-4
-2
-0.76
0
2
4
x

Use a Z Table

- Can't calculate the area directly (unless you like calculus)
- Use Z tables, printed in the back of text books and available online

$$\Phi(Z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{Z} e^{-\frac{t^2}{2}} dt$$

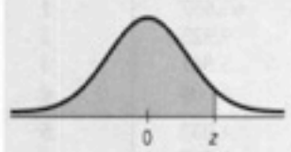
Table Z
Areas under the standard Normal curve



z	Second decimal place in z									
	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
−3.9	0.0000†	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
−3.8	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
−3.7	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
−3.6	0.0002	0.0002	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
−3.5	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002
−3.4	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0002
−3.3	0.0005	0.0005	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0003
−3.2	0.0007	0.0007	0.0006	0.0006	0.0006	0.0006	0.0006	0.0005	0.0005	0.0005
−3.1	0.0010	0.0009	0.0009	0.0009	0.0008	0.0008	0.0008	0.0008	0.0007	0.0007
−3.0	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011	0.0010	0.0010
−2.9	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
−2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
−2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
−2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
−2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
−2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
−2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
−2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
−2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
−2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
−1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
−1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
−1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
−1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
−1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
−1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
−1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
−1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
−1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
−1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
−0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
−0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
−0.7	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
−0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
−0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
−0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
−0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
−0.2	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
−0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
−0.0	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641

†For $z \leq -3.90$, the areas are 0.0000 to four decimal places.

Table Z (cont.)
Areas under the standard Normal curve

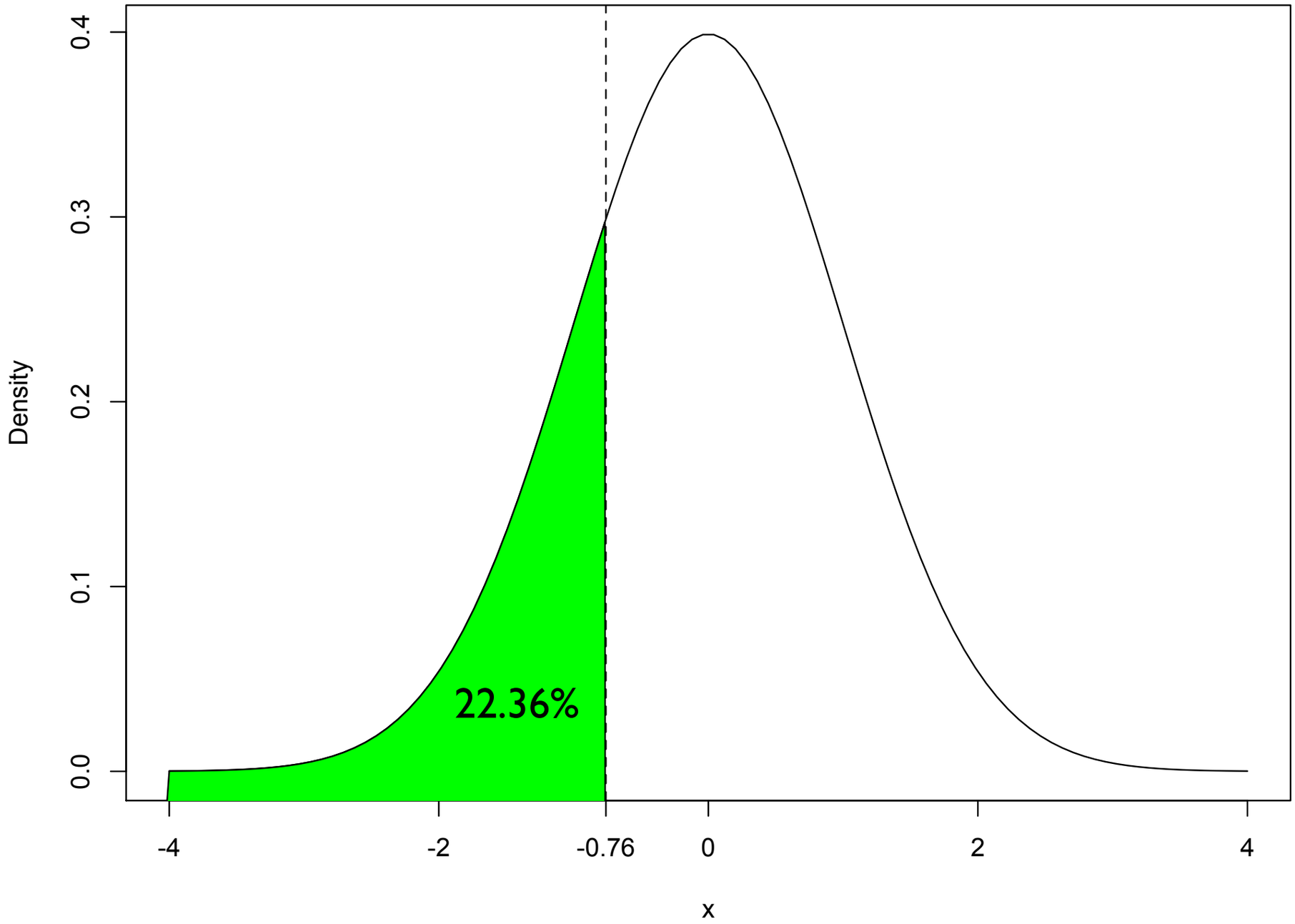


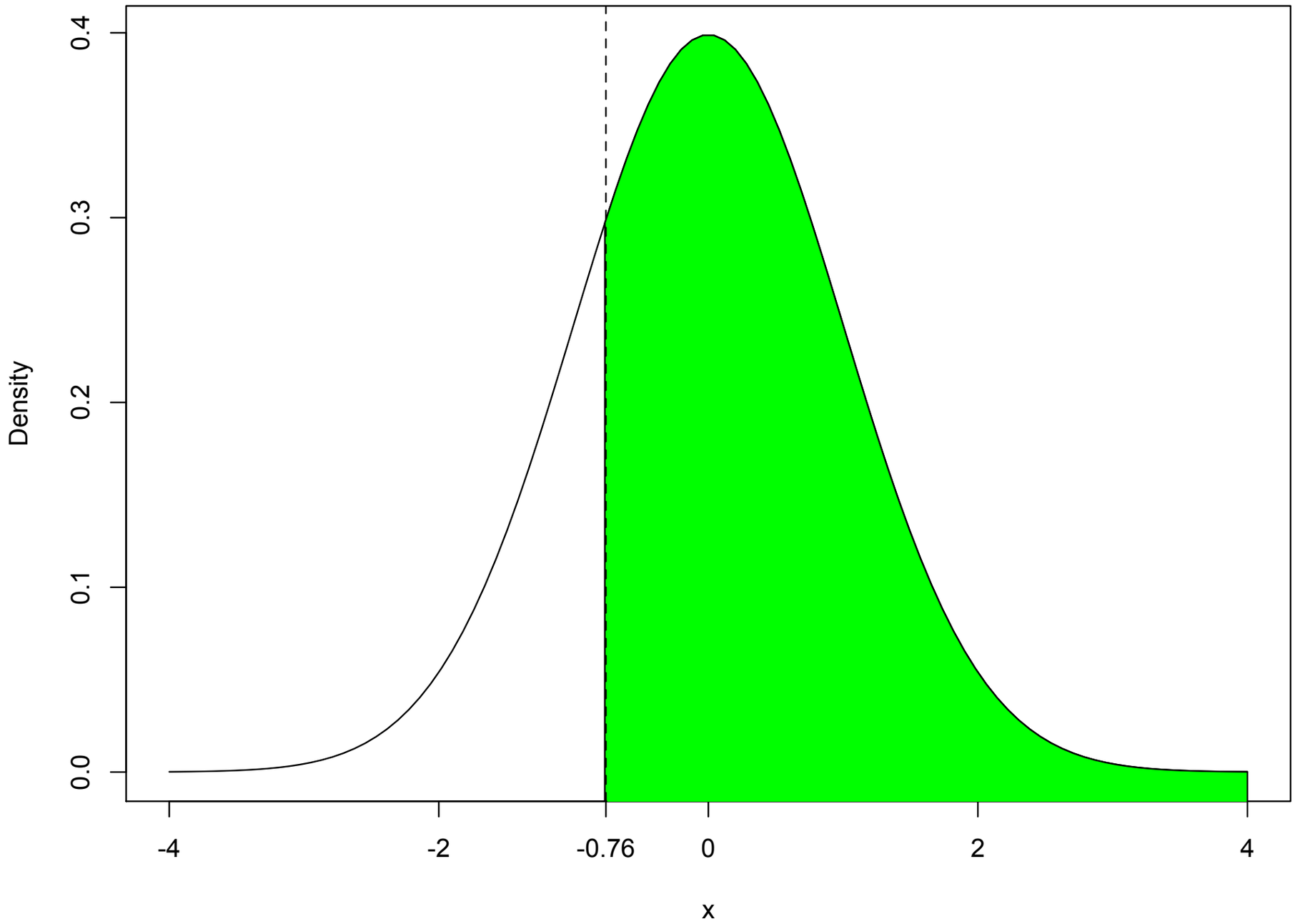
z	Second decimal place in z									
	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998
3.5	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998
3.6	0.9998	0.9998	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.7	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.8	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.9	1.0000†									

†For $z \geq 3.90$, the areas are 1.0000 to four decimal places.

Second decimal place in z										
0.09	*0.08*	*0.07*	*0.06*	*0.05*	*0.04*	*0.03*	*0.02*	*0.01*	*0.00*	z
0.1611	0.1635	0.1660	0.1685	0.1711	0.1736	0.1762	0.1788	0.1814	0.1841	-0.9
0.1867	0.1894	0.1922	0.1949	0.1977	0.2005	0.2033	0.2061	0.2090	0.2119	-0.8
0.2148	0.2177	0.2206	0.2236	0.2266	0.2296	0.2327	0.2358	0.2389	0.2420	-0.7
0.2451	0.2483	0.2514	0.2546	0.2578	0.2611	0.2643	0.2676	0.2709	0.2743	-0.6
0.2776	0.2810	0.2843	0.2877	0.2912	0.2946	0.2981	0.3015	0.3050	0.3085	-0.5

$$P(Z \leq -0.76) = 0.2236$$


Density
0.0
0.1
0.2
0.3
0.4
22.36%
-4
-2
-0.76
0
2
4
x


Density
0.0
0.1
0.2
0.3
0.4
-4
-2
-0.76
0
2
4
x

Z scores

- To find the proportion above a number, we can subtract the proportion below from 1

$$P(Z \geq -0.76) = 1 - 0.2236 = 0.7764$$

Z scores

- Can also think of these values as probabilities
- Say you have a set of observations with some mean and standard deviation, what is the probability of picking a value greater than some number?

Sample Question

- Given a set of normally distributed observations with mean of 25 and standard deviation of 5, what is the probability of randomly choosing a number less than 10?

$$\bar{X} = 25 \qquad s = 5 \qquad X_i = 10$$

$$Z = \frac{10 - 25}{5} = \frac{-15}{5} = -3$$

0.00	z
0.0000†	−3.9
0.0001	−3.8
0.0001	−3.7
0.0002	−3.6
0.0002	−3.5
0.0003	−3.4
0.0005	−3.3
0.0007	−3.2
0.0010	−3.1
0.0013	−3.0

$$P(X \leq 10) = P(Z \leq -3) = 0.0013$$

The probability of randomly selecting a number less than 10 is 0.0013 or 0.13%

Inference

- Using this method, we can make inferences about a population using a random sample of the population

Confidence Intervals

Confidence Intervals

- When we take a sample, we can calculate the sample's mean
  - This is a point estimate of the population mean
- But what is the true population mean?
- Construct confidence intervals to estimate

Confidence Intervals

- Contain a point estimate of the population mean, plus or minus some margin of error
- The size of the margin of error is determined by how confident we want to be

Confidence Intervals

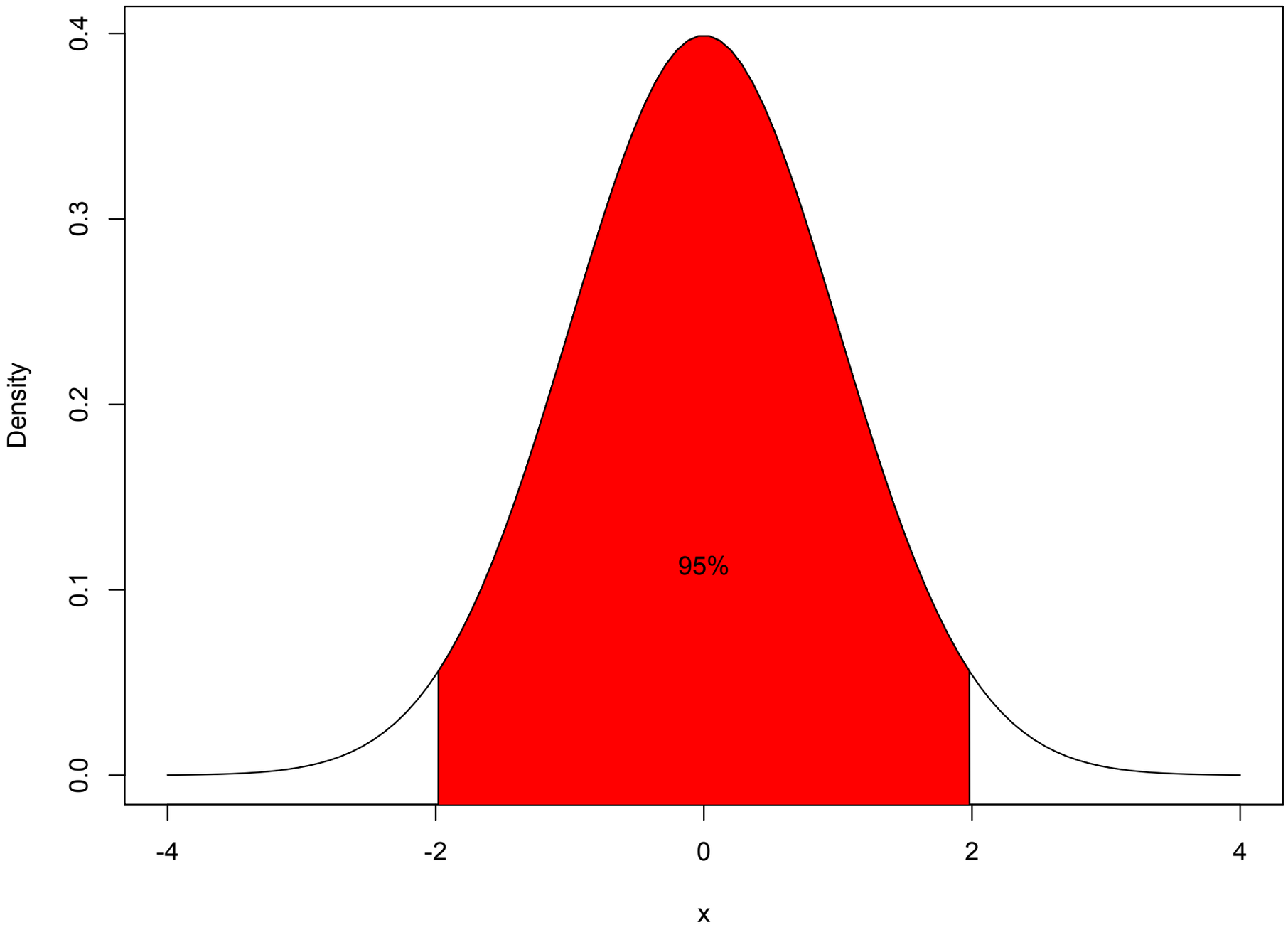
$$\bar{X} \pm Z(\sigma_{\bar{X}})$$

But we don't know the standard error, so we estimate it using the sample's standard deviation

$$\sigma_{\bar{X}} = \frac{s}{\sqrt{n}}$$

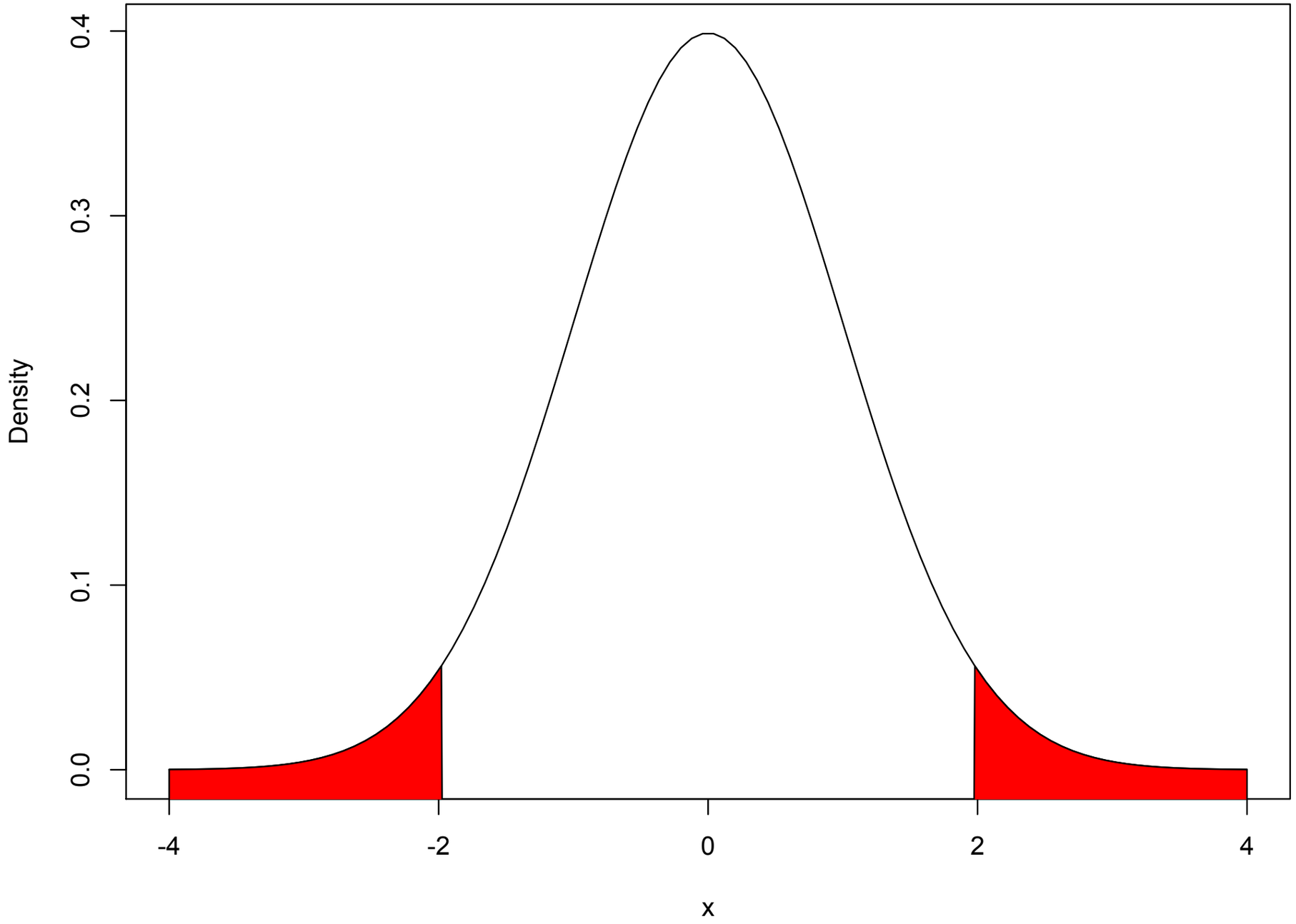
$$\bar{X} \pm Z\left(\frac{s}{\sqrt{n}}\right)$$

Recall:



Confidence Intervals

- So if we want to be 95% confident that the population mean lies within the confidence interval, we should use a Z score close to 2
- How do we figure out which Z we should use?



If we want a 95% confidence interval, we will leave 5% outside, so each tail will have 2.5% or 0.0250

What Z score corresponds with 0.0250 in the lower tail?

Second decimal place in z										
0.09	*0.08*	*0.07*	*0.06*	*0.05*	*0.04*	*0.03*	*0.02*	*0.01*	*0.00*	*z*
0.0233	0.0239	0.0244	0.0250	0.0256	0.0262	0.0268	0.0274	0.0281	0.0287	−1.9
0.0294	0.0301	0.0307	0.0314	0.0322	0.0329	0.0336	0.0344	0.0351	0.0359	−1.8
0.0367	0.0375	0.0384	0.0392	0.0401	0.0409	0.0418	0.0427	0.0436	0.0446	−1.7
0.0455	0.0465	0.0475	0.0485	0.0495	0.0505	0.0516	0.0526	0.0537	0.0548	−1.6
0.0559	0.0571	0.0582	0.0594	0.0606	0.0618	0.0630	0.0643	0.0655	0.0668	−1.5

The Z score that corresponds with 0.025 in the lower tail is -1.96. This is the Z score we will use for a 95% confidence interval

Given the following statistics, construct a 95% confidence interval for the population mean

$$\bar{X} = 25 \qquad s = 10 \qquad n = 100$$

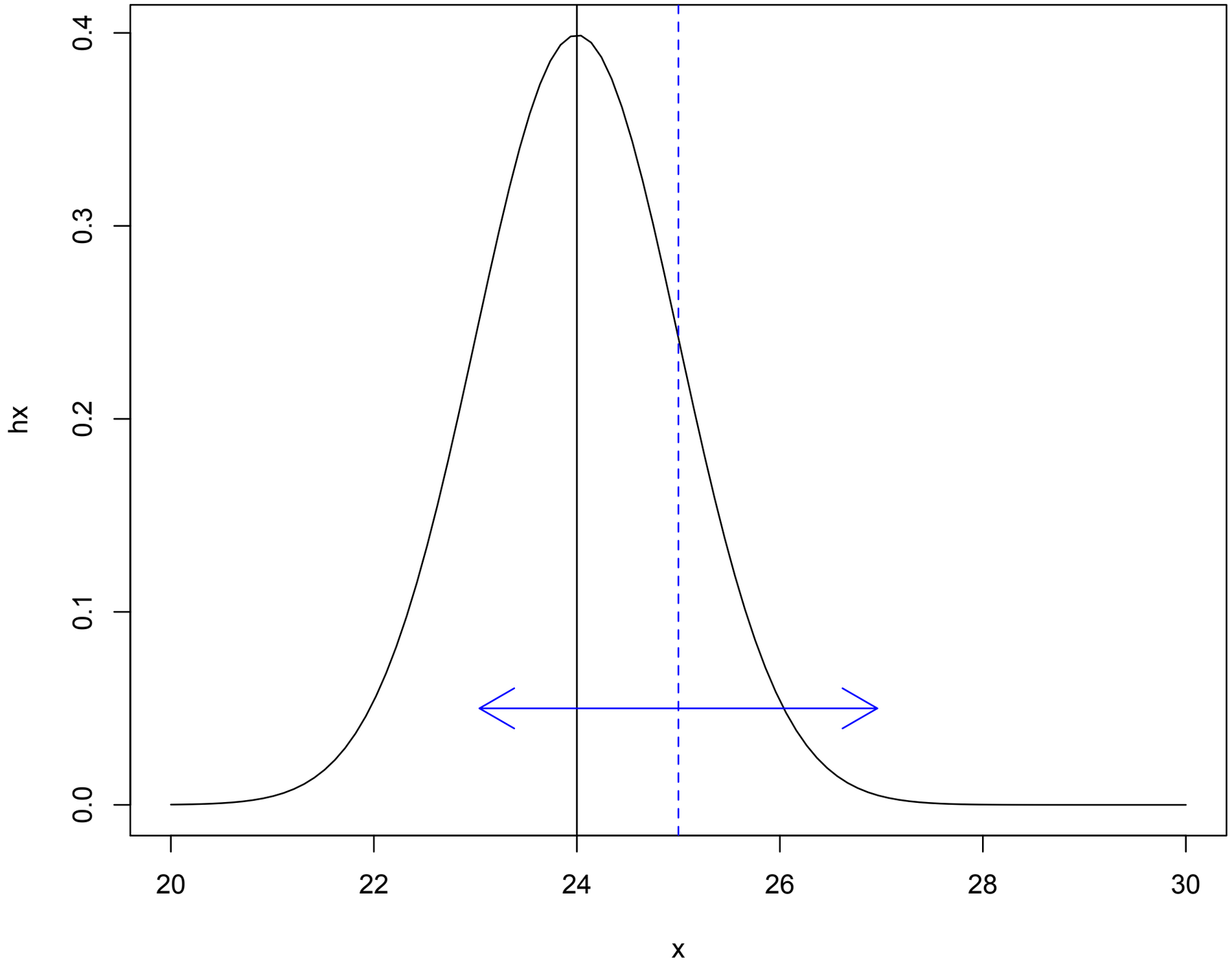
$$\bar{X} \pm Z\left(\frac{s}{\sqrt{n}}\right)$$

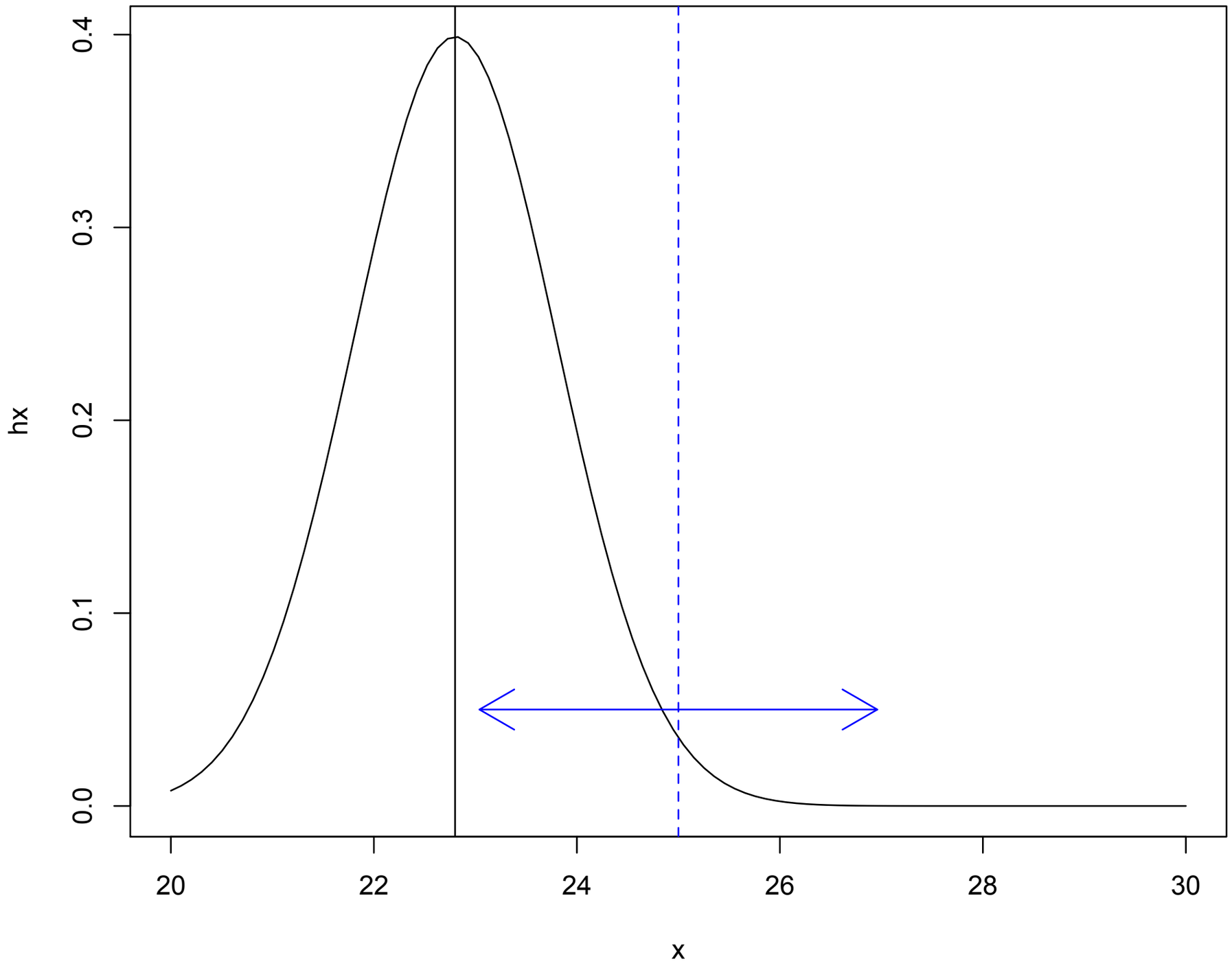
$$25 \pm 1.96\left(\frac{10}{\sqrt{10}}\right)$$

$$25 \pm 1.96(1)$$

$$\boxed{25 \pm 1.96}$$

We can be 95% confidence that the population mean is between 23.04 and 26.96.


hx
0.0
0.1
0.2
0.3
0.4
20
22
24
26
28
30
x

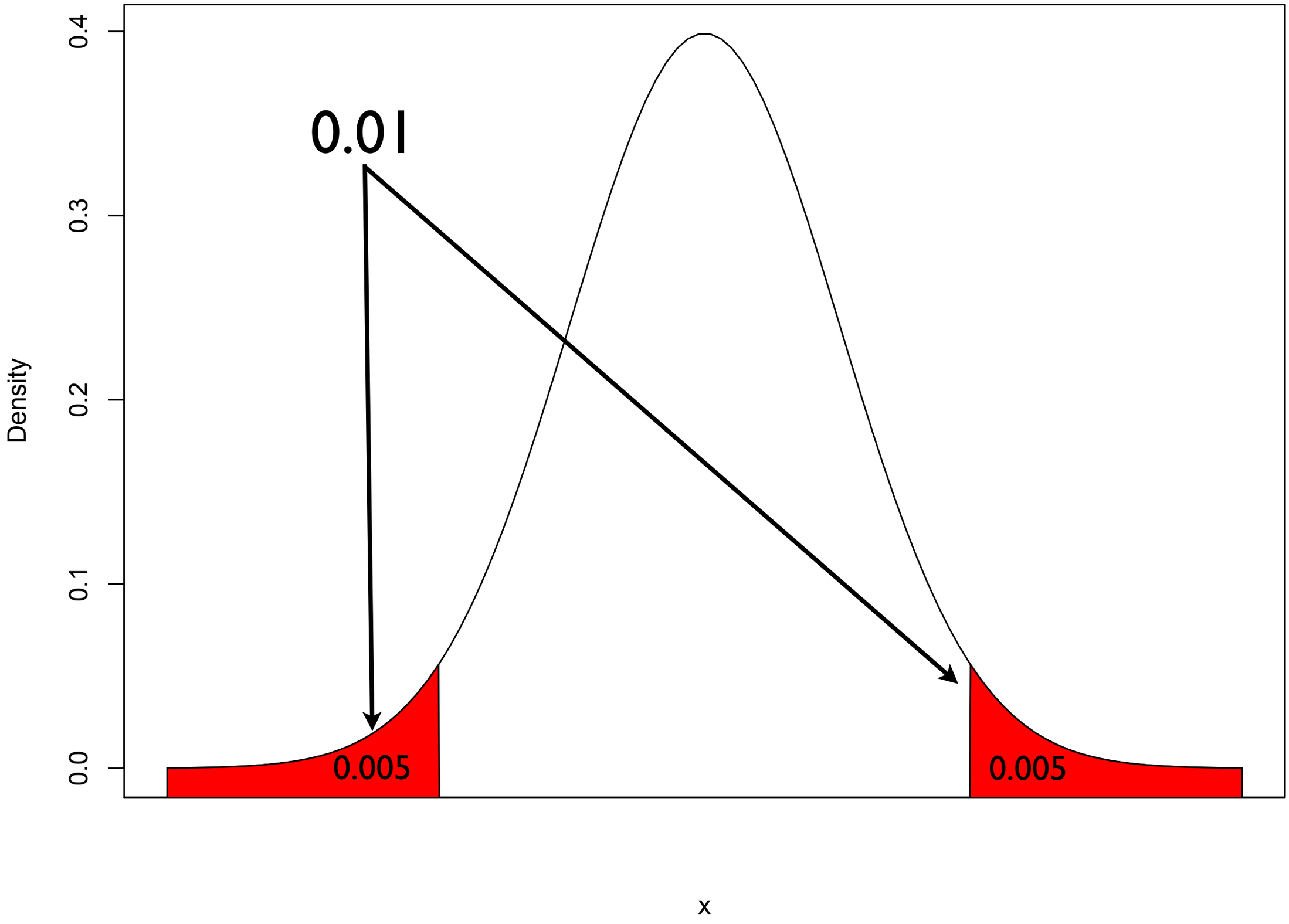

hx
0.0
0.1
0.2
0.3
0.4
20
22
24
26
28
30
x

Confidence Intervals

- A 95% confidence interval means that there is a 5% chance that we are wrong
- This is sometimes referred to as our alpha
- Alpha is the chance of making a type I error (more on these later)

Confidence Intervals

- What if we wanted to be more certain? We wanted a lower alpha?
- We can construct a confidence interval with higher percentages - 99% for example


0.01
0.005
0.005
Density
0.0
0.1
0.2
0.3
0.4
x

Second decimal place in z										
0.09	*0.08*	*0.07*	*0.06*	*0.05*	*0.04*	*0.03*	*0.02*	*0.01*	*0.00*	z
0.0014	0.0014	0.0015	0.0015	0.0016	0.0016	0.0017	0.0018	0.0018	0.0019	−2.9
0.0019	0.0020	0.0021	0.0021	0.0022	0.0023	0.0023	0.0024	0.0025	0.0026	−2.8
0.0026	0.0027	0.0028	0.0029	0.0030	0.0031	0.0032	0.0033	0.0034	0.0035	−2.7
0.0036	0.0037	0.0038	0.0039	0.0040	0.0041	0.0043	0.0044	0.0045	0.0047	−2.6
0.0048	0.0049	0.0051	0.0052	0.0054	0.0055	0.0057	0.0059	0.0060	0.0062	−2.5

0.005 lies somewhere between -2.58 and -2.57

We could use -2.575, or just -2.58

Given the following statistics, construct a 95% confidence interval for the population mean

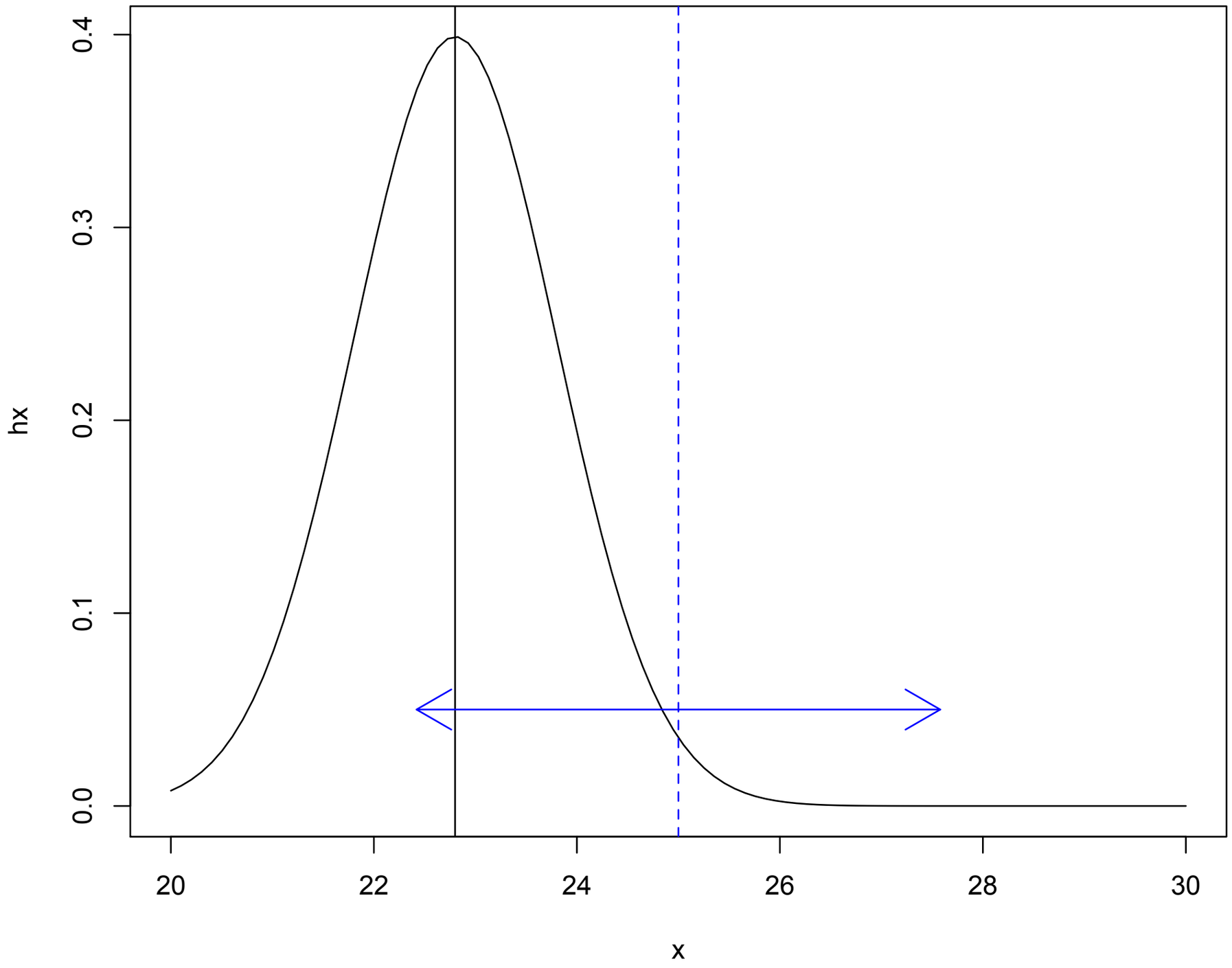
$$\bar{X} = 25 \qquad s = 10 \qquad n = 100$$

$$\bar{X} \pm Z\left(\frac{s}{\sqrt{n}}\right)$$

$$25 \pm 2.58\left(\frac{10}{\sqrt{10}}\right)$$

$$25 \pm 2.58$$

We can be 95% confidence that the population mean is between 22.42 and 27.58.


hx
0.0
0.1
0.2
0.3
0.4
20
22
24
26
28
30
x

Confidence Intervals

- Notice the confidence interval got bigger when we increased our confidence level
- Since we want there to be less chance of being wrong, our interval has to encompass more potential values

Practice!

- You interview 150 people and ask their age. The mean age was 45 with a standard deviation of 18. Construct a 95% confidence interval for the population mean age.

$$\bar{X} = 45 \qquad s = 18 \qquad n = 150$$

$$\bar{X} \pm Z\left(\frac{s}{\sqrt{n}}\right)$$

$$45 \pm 1.96\left(\frac{18}{\sqrt{150}}\right)$$

$$45 \pm 1.96(1.47)$$

$$45 \pm 2.89$$

We are 95% confident that the population mean age is between 42.11 years and 47.89 years old.

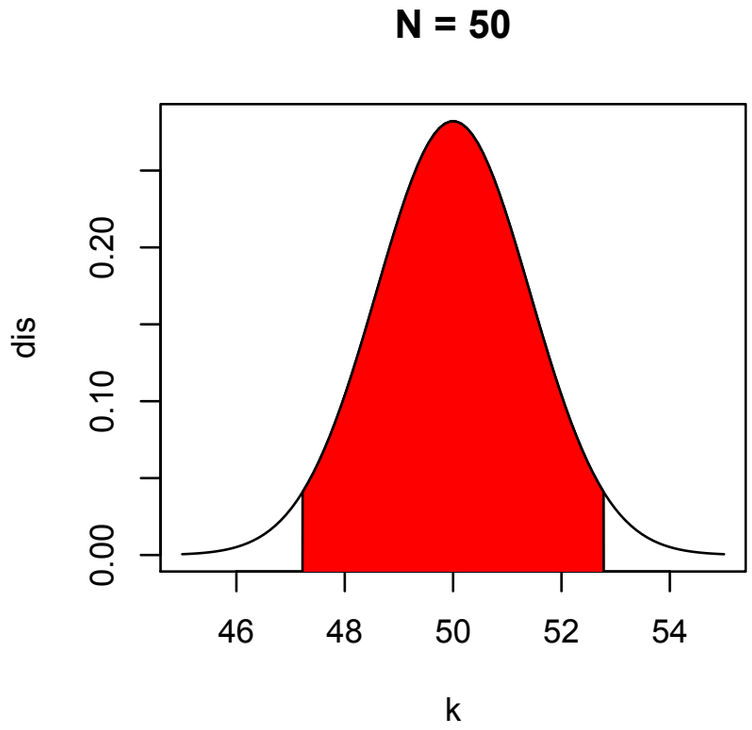
Sample Size

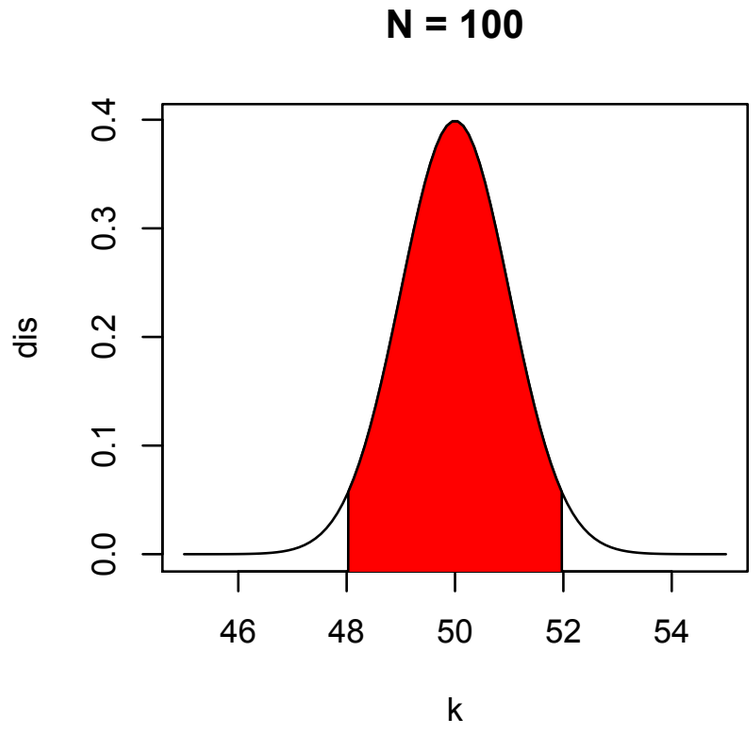
Sample Size

- Notice that the sample size is included in the calculation of the standard error:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

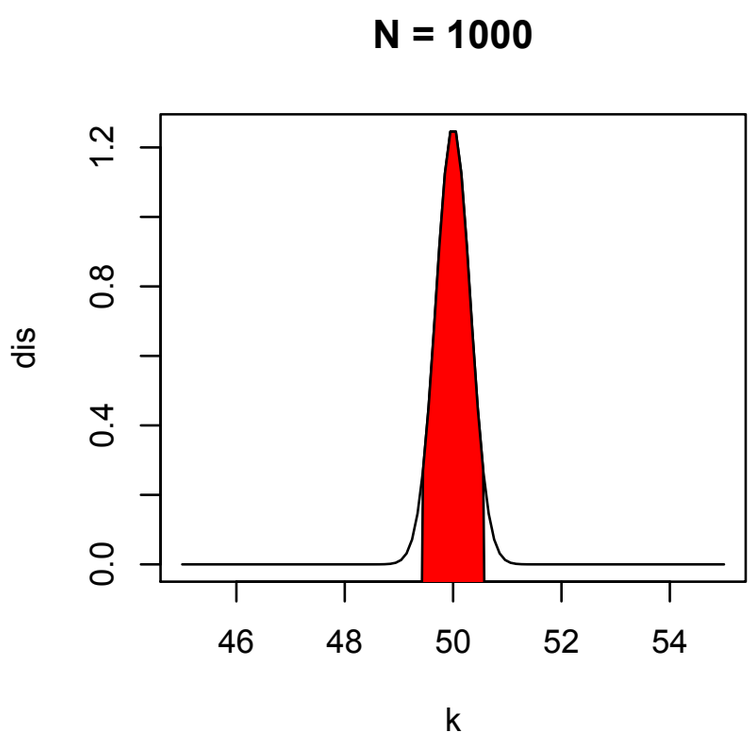
- So larger sample sizes will have smaller standard errors, and therefore narrower confidence intervals

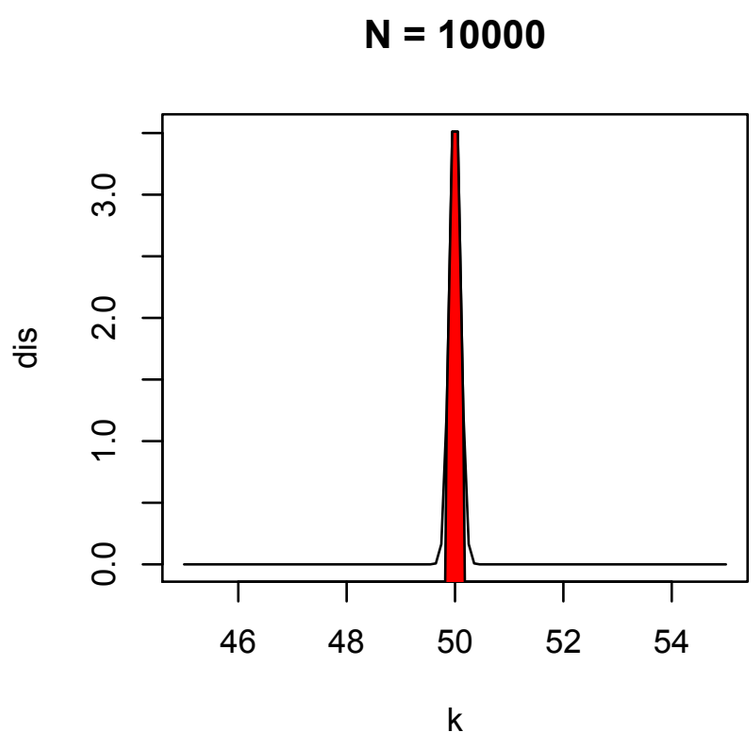

N = 50
dis
0.00
0.10
0.20
46
48
50
52
54
k


N = 100
dis
0.0
0.1
0.2
0.3
0.4
46
48
50
52
54
k

$$\bar{x} = 50$$

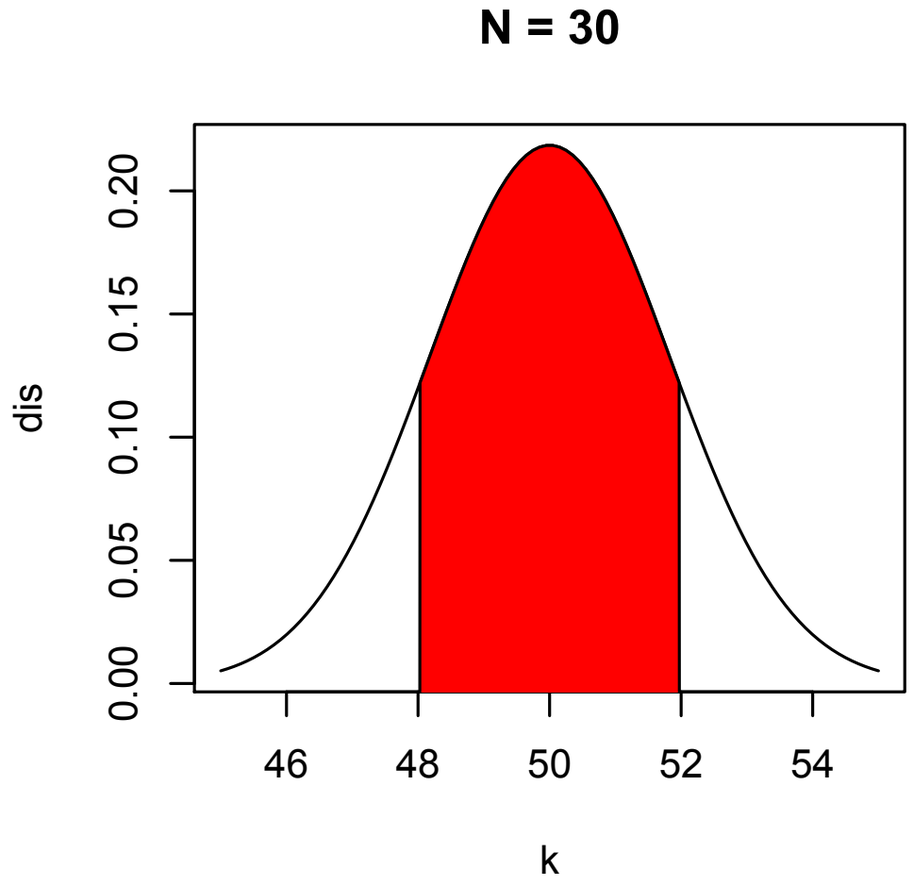
$$s = 10$$

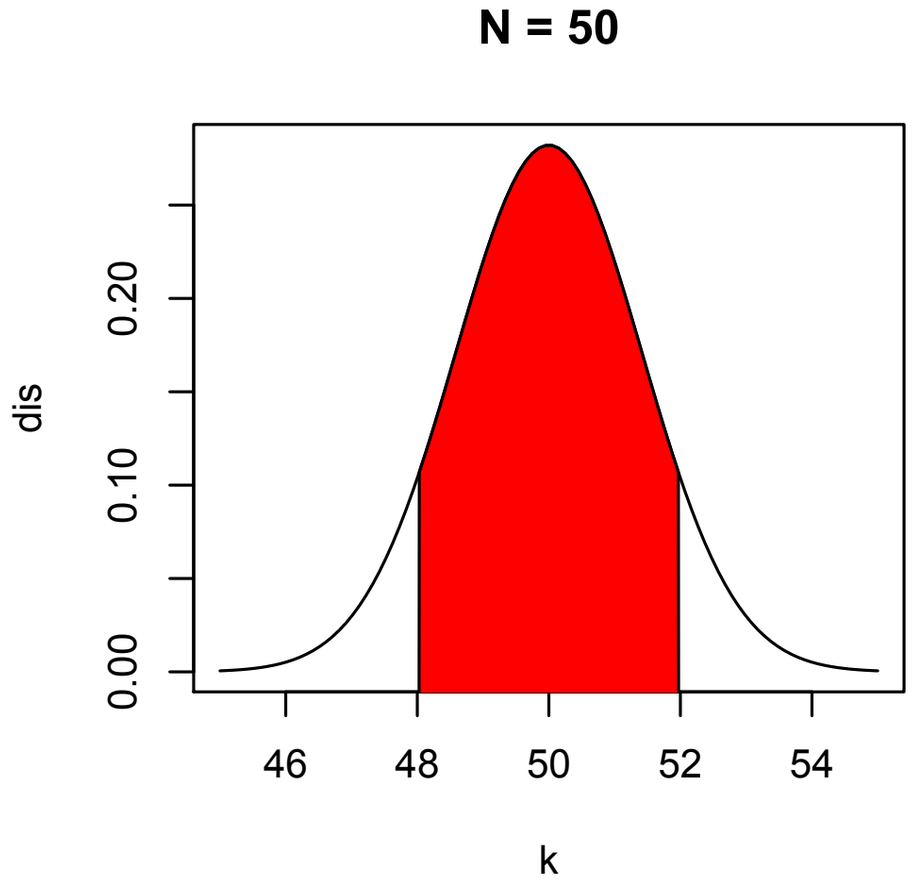

N = 1000
dis
0.0
0.4
0.8
1.2
46
48
50
52
54
k

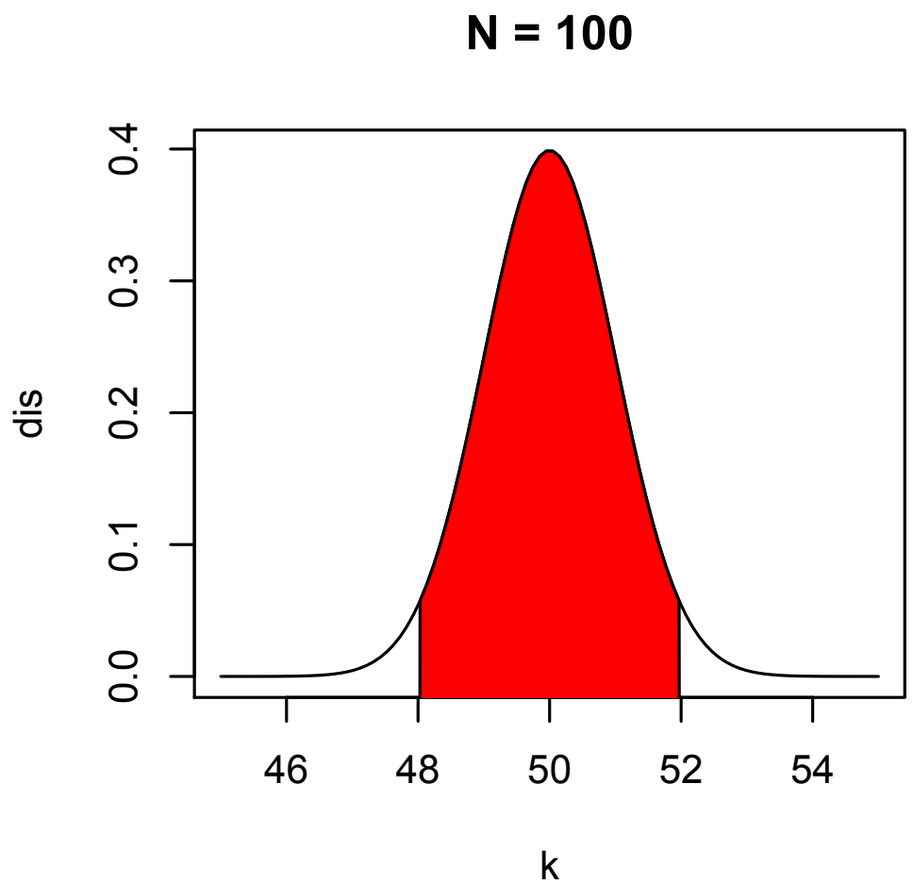

N = 10000
dis
0.0
1.0
2.0
3.0
46
48
50
52
54
k

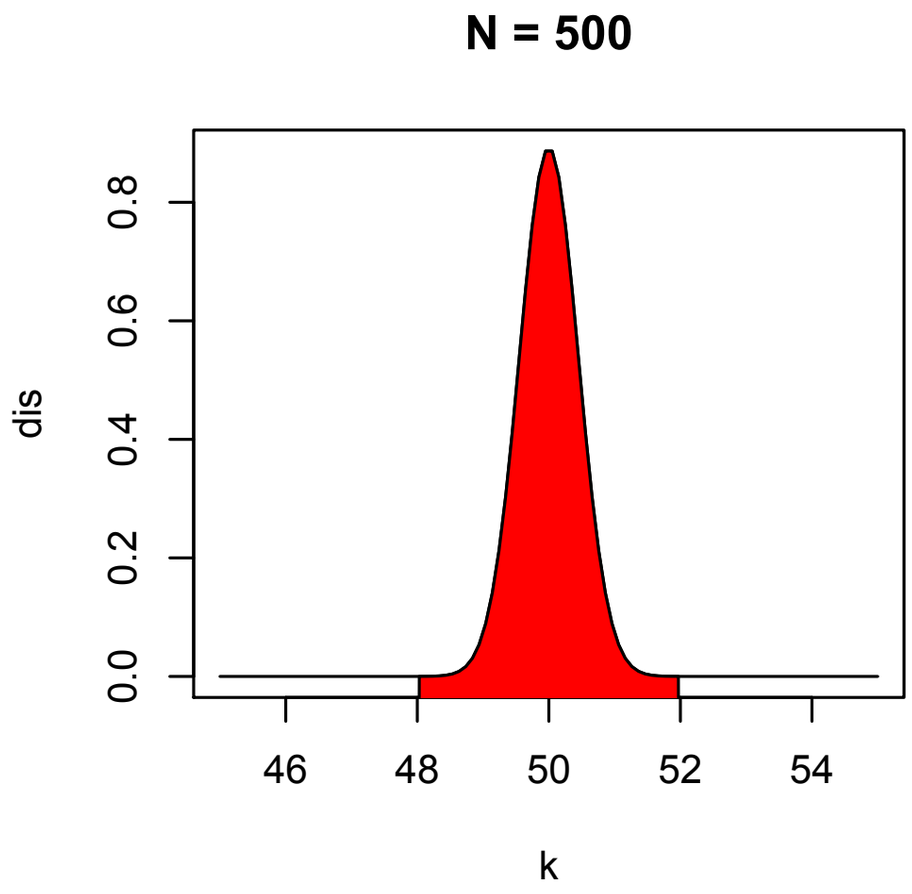
If our confidence interval was fixed at 48 to 52, what would the corresponding Z scores and confidence levels be for increasing sample sizes?

n	Z	Confidence level
30	1.095	72.7%
50	1.414	84.3%
100	2	95.4%
500	4.472	99.99923%


N = 30
dis
0.00
0.05
0.10
0.15
0.20
46
48
50
52
54
k


N = 50
dis
0.00
0.10
0.20
46
48
50
52
54
k


N = 100
dis
0.0
0.1
0.2
0.3
0.4
46
48
50
52
54
k


N = 500
dis
0.0
0.2
0.4
0.6
0.8
46
48
50
52
54
k

Sample Size

- So larger sample sizes give you more statistical power - you can be more confident that the population mean is closer to your sample mean.

Proportions

Proportions

- So far we've been talking about sample and population means - when our variable is interval or ratio scale
- What about when our variable is categorical?

Proportions

- What is the proportion of male-identified students in a classroom?
- What is the proportion of college graduates in a firm?
- Proportions can be treated like a mean
  - In fact, if you code categorical variables as 0/1, the mean of the variables is the proportion of the 1 category

Proportions

- Today we are discussing proportions that results from Bernoulli trials
  - Only two outcomes
  - same probability for success
  - each trial is independent of other trials

Proportions

- Such trials produce binomial distributions
- For large enough n, binomial distributions approximate normal distributions
  - n*p > 5 and n*(1-p) > 5

Notation

$$P(1) = \pi$$

$$P(0) = 1 - \pi$$

$$\pi = \frac{X}{N} = \frac{\text{successes}}{N}$$

$$\text{sample proportion} = \hat{\pi}$$

Population
Standard Deviation

$$\sigma = \sqrt{\pi(1-\pi)}$$

Standard Error of the
Sample Proportion

$$\sigma_{\hat{\pi}} = \frac{\sigma}{\sqrt{N}} = \sqrt{\frac{\pi(1-\pi)}{N}}$$

Confidence Intervals

- We can calculate confidence intervals for a proportion just like we did with a mean

$$\hat{\pi} \pm Z\sigma_{\hat{\pi}}$$

$$\hat{\pi} \pm Z\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}$$

Sample Problem

- You are interested in exercise habits of college students. You randomly sample 100 UCI students and ask them if they go to the gym at least once a week. 74 students respond yes. Construct a 95% confidence interval for the proportion of UCI students who go to the gym regularly.

$$n = 100 \qquad x = 74$$

$$\hat{\pi} = \frac{74}{100} = 0.74$$

$$\hat{\pi} \pm Z\sigma_{\hat{\pi}}$$

$$0.74 \pm 1.96(0.0439)$$

$$0.74 \pm 0.086$$

$$\sigma = \sqrt{\hat{\pi}(1-\hat{\pi})}$$
$$= \sqrt{0.74(1-0.74)}$$
$$= \sqrt{0.1924} = 0.4386$$

$$\sigma_{\hat{\pi}} = \frac{\sigma}{\sqrt{n}} = \frac{0.4386}{\sqrt{100}}$$
$$= 0.0439$$

We are 95% confident that the proportion of UCI students who exercise regularly is between 0.654 and 0.826

Practice!

- A survey company randomly samples 500 Americans and finds that 310 of them prefer cats to dogs. Construct a 95% confidence interval for the proportion of cat lovers in the US.

$$n = 500 \qquad x = 310$$

$$\hat{\pi} = \frac{310}{500} = 0.62$$

$$\sigma = \sqrt{\hat{\pi}(1-\hat{\pi})}$$
$$= \sqrt{0.62(1-0.62)}$$
$$= 0.4854$$

$$\hat{\pi} \pm Z\sigma_{\hat{\pi}}$$

$$0.62 \pm 1.96(0.0217)$$

$$\sigma_{\hat{\pi}} = \frac{\sigma}{\sqrt{n}} = \frac{0.4854}{\sqrt{500}}$$
$$= 0.0217$$

$$0.62 \pm 0.0425$$

We are 95% confident that the proportion of cat lovers in the US is between 0.5775 and 0.6625

Recap

- Inferential statistics lets us make inferences about a population based on a sample
- Populations have parameters, samples have statistics

Recap

- Lots of possible samples from one population
- The distribution of sample statistics from all the possible samples of a population is called the sampling distribution

Recap

Sampling Distribution Mean

$$\mu_{\bar{X}} = \frac{\sum \bar{X}}{N}$$

Standard Error

$$\sigma_{\bar{X}} = \sqrt{\frac{\sum (X - \mu_{\bar{X}})^2}{N}}$$

Recap

- Z scores are standardized deviation scores

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

Recap

- We can use the sampling distribution to construct confidence intervals for population means and proportions

$$\bar{X} \pm Z(\sigma_{\bar{X}})$$

$$\hat{\pi} \pm Z\sigma_{\hat{\pi}}$$

Student's t-distribution

Assumption of Normality

- So far we've been assuming that distributions are normal
- But sometimes we can't assume that
- Notably, when our samples are small

Small Samples

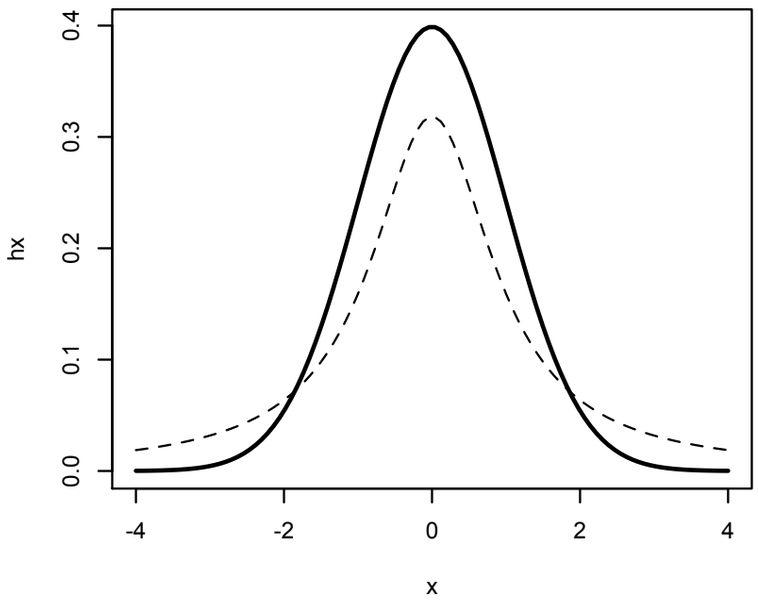
- Small samples tend to have more error - more likely for sample statistics to be different than population parameters
- So confidence intervals need to be wider to account for this additional error

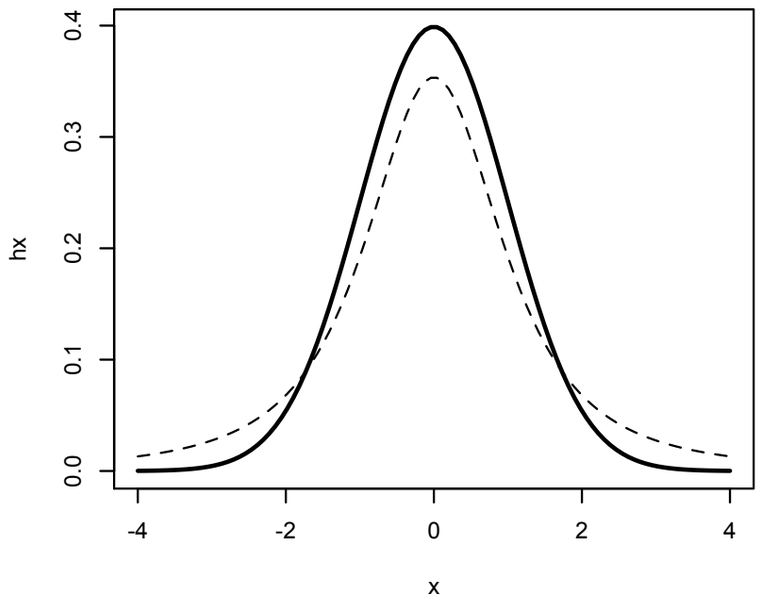
t-distribution

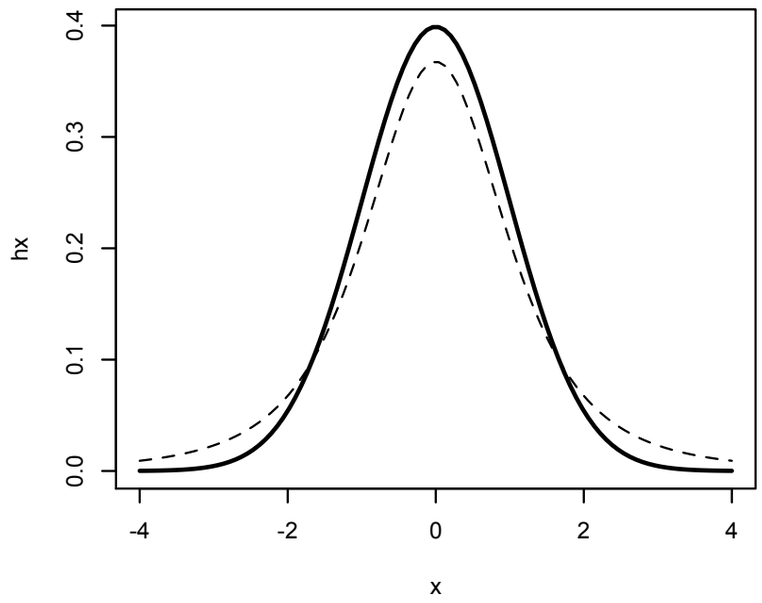
- The t-distribution is used when samples are small
  - Has an additional assumption - that the population is normally distributed
- The exact shape of the distribution depends on your degrees of freedom

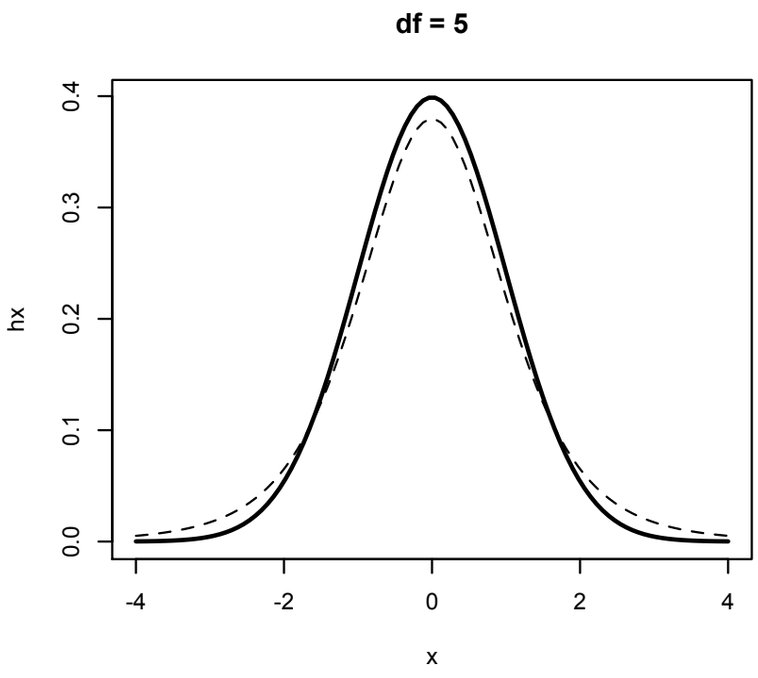
Degrees of Freedom ν

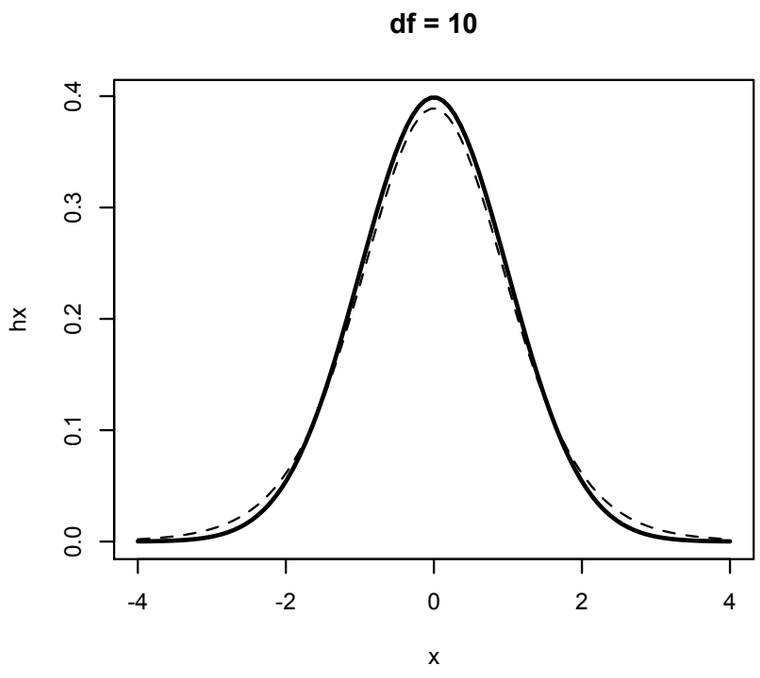
- In general, degrees of freedom refer to the number of values in the final calculation of a statistic that are free to vary
- This is typically the number of observations minus the number of statistics or parameters calculated
- For one mean/proportion instances, like we are covering in this course, degrees of freedom is n-1

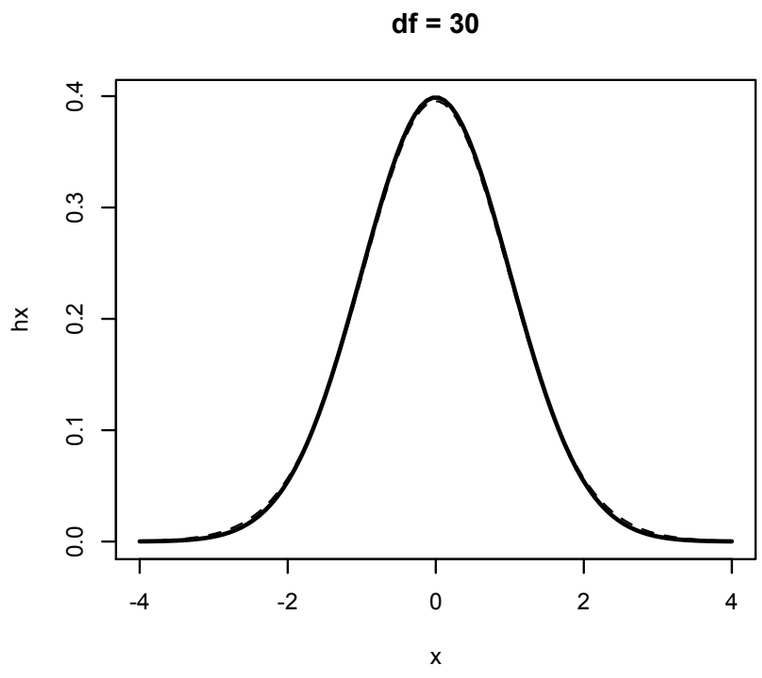

df = 1
hx
x
0.0
0.1
0.2
0.3
0.4
-4
-2
0
2
4


df = 2
hx
x
0.0
0.1
0.2
0.3
0.4
-4
-2
0
2
4


df = 3
hx
x
0.0
0.1
0.2
0.3
0.4
-4
-2
0
2
4


df = 5
hx
x
0.0
0.1
0.2
0.3
0.4
-4
-2
0
2
4


df = 10
hx
x
0.0
0.1
0.2
0.3
0.4
-4
-2
0
2
4


df = 30
hx
x
0.0
0.1
0.2
0.3
0.4
-4
-2
0
2
4

t-distribution

- Notice that as degrees of freedom increases, the t-distribution becomes the normal distribution
- By n=30, t scores are approximately z scores

Sample Problem

- You survey 5 people on the street and find that their mean income is $25,000 per year, with a standard deviation of $10,000. Construct a 95% confidence interval for the population mean income.

$$\nu = 5 - 1 = 4$$

	Confidence Level					
	80%	90%	95%	98%	99%	99.8%
	Right-Tail Probability					
df	$t_{.100}$	$t_{.050}$	$t_{.025}$	$t_{.010}$	$t_{.005}$	$t_{.001}$
1	3.078	6.314	12.706	31.821	63.656	318.289
2	1.886	2.920	4.303	6.965	9.925	22.328
3	1.638	2.353	3.182	4.541	5.841	10.214
4	1.533	2.132	2.776	3.747	4.604	7.173
5	1.476	2.015	2.571	3.365	4.032	5.894
6	1.440	1.943	2.447	3.143	3.707	5.208
7	1.415	1.895	2.365	2.998	3.499	4.785
8	1.397	1.860	2.306	2.896	3.355	4.501
9	1.383	1.833	2.262	2.821	3.250	4.297
10	1.372	1.812	2.228	2.764	3.169	4.144
11	1.363	1.796	2.201	2.718	3.106	4.025
12	1.356	1.782	2.179	2.681	3.055	3.930
13	1.350	1.771	2.160	2.650	3.012	3.852
14	1.345	1.761	2.145	2.624	2.977	3.787
15	1.341	1.753	2.131	2.602	2.947	3.733
16	1.337	1.746	2.120	2.583	2.921	3.686
17	1.333	1.740	2.110	2.567	2.898	3.646
18	1.330	1.734	2.101	2.552	2.878	3.611

$$n = 5 \quad \bar{x} = 25000 \quad s = 10000$$

$$t = 2.776$$

$$\bar{x} \pm t\sigma_{\bar{x}} \qquad \bar{x} \pm t\left(\frac{s}{\sqrt{n}}\right)$$

$$25000 \pm 2.776\left(\frac{10000}{\sqrt{5}}\right)$$

$$25000 \pm 2.776(4472.14)$$

$$25000 \pm 12414.65$$

We can be 95% confident that the population mean income is between $12,585.35 and $37,414.65 per year.

Practice!

- You survey 8 faculty and discover their mean number of publications is 20 with a standard deviation of 5. Construct a 99% confidence interval for the population mean number of publications.

	Confidence Level					
	80%	90%	95%	98%	99%	99.8%
	Right-Tail Probability					
df	$t_{.100}$	$t_{.050}$	$t_{.025}$	$t_{.010}$	$t_{.005}$	$t_{.001}$
1	3.078	6.314	12.706	31.821	63.656	318.289
2	1.886	2.920	4.303	6.965	9.925	22.328
3	1.638	2.353	3.182	4.541	5.841	10.214
4	1.533	2.132	2.776	3.747	4.604	7.173
5	1.476	2.015	2.571	3.365	4.032	5.894
6	1.440	1.943	2.447	3.143	3.707	5.208
7	1.415	1.895	2.365	2.998	3.499	4.785
8	1.397	1.860	2.306	2.896	3.355	4.501
9	1.383	1.833	2.262	2.821	3.250	4.297
10	1.372	1.812	2.228	2.764	3.169	4.144
11	1.363	1.796	2.201	2.718	3.106	4.025
12	1.356	1.782	2.179	2.681	3.055	3.930
13	1.350	1.771	2.160	2.650	3.012	3.852
14	1.345	1.761	2.145	2.624	2.977	3.787
15	1.341	1.753	2.131	2.602	2.947	3.733
16	1.337	1.746	2.120	2.583	2.921	3.686
17	1.333	1.740	2.110	2.567	2.898	3.646
18	1.330	1.734	2.101	2.552	2.878	3.611
19	1.328	1.729	2.093	2.539	2.861	3.579

$$n = 8 \quad \bar{x} = 20 \quad s = 5 \quad t = 3.499$$

$$\bar{x} \pm t\sigma_{\bar{x}} \qquad \bar{x} \pm t\left(\frac{s}{\sqrt{n}}\right)$$

$$20 \pm 3.499\left(\frac{5}{\sqrt{8}}\right)$$

$$20 \pm 3.499\,(1.768)$$

$$20 \pm 6.185$$

We can be 95% confident that the population mean number of publications is between 13.815 and 26.185 publications.