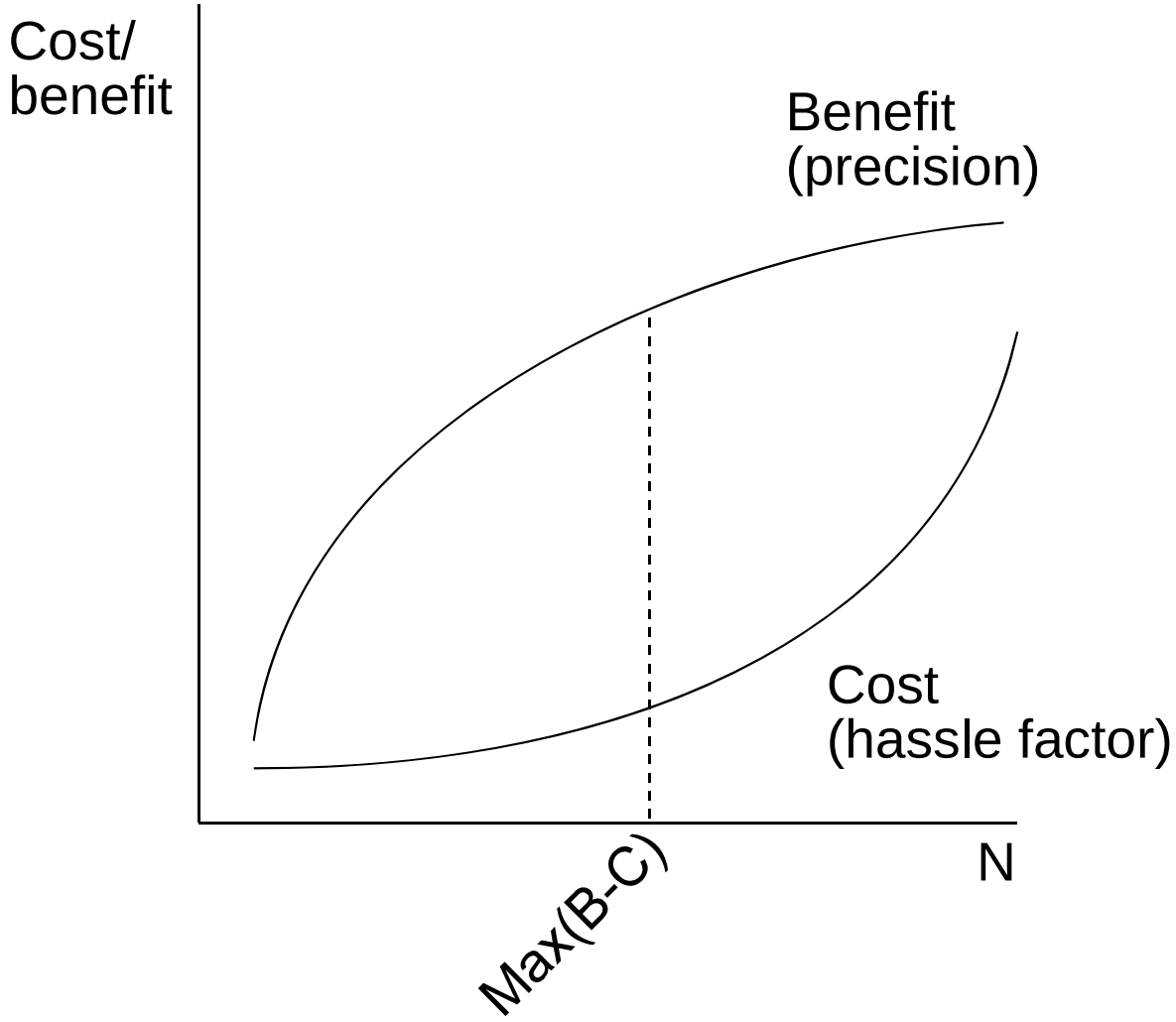
Sampling and Inference

The Quality of Data and Measures

2013

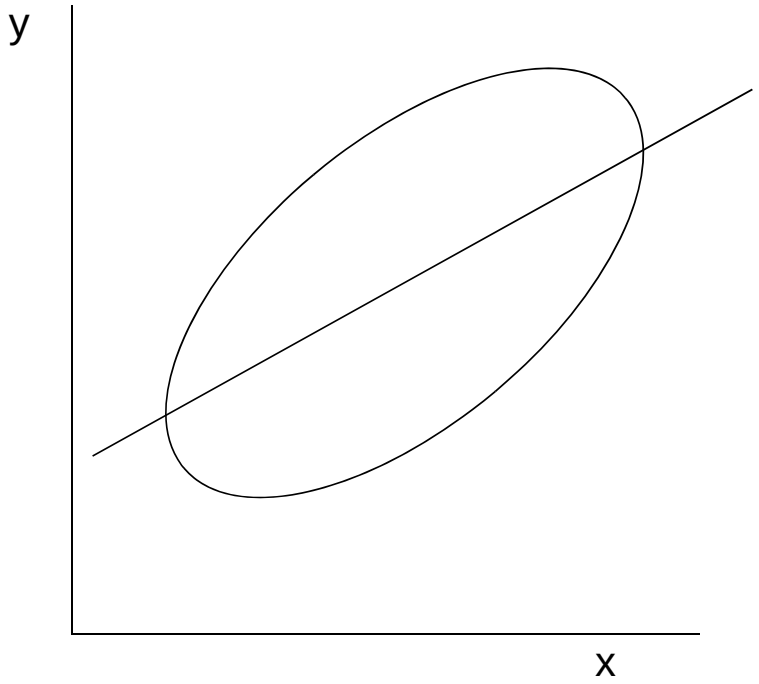
Why do we sample?

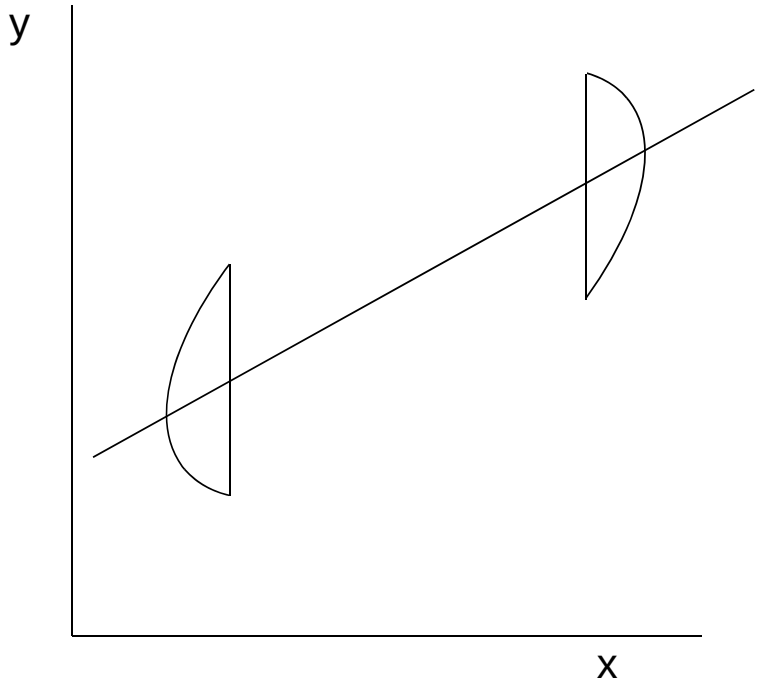


Effects of samples

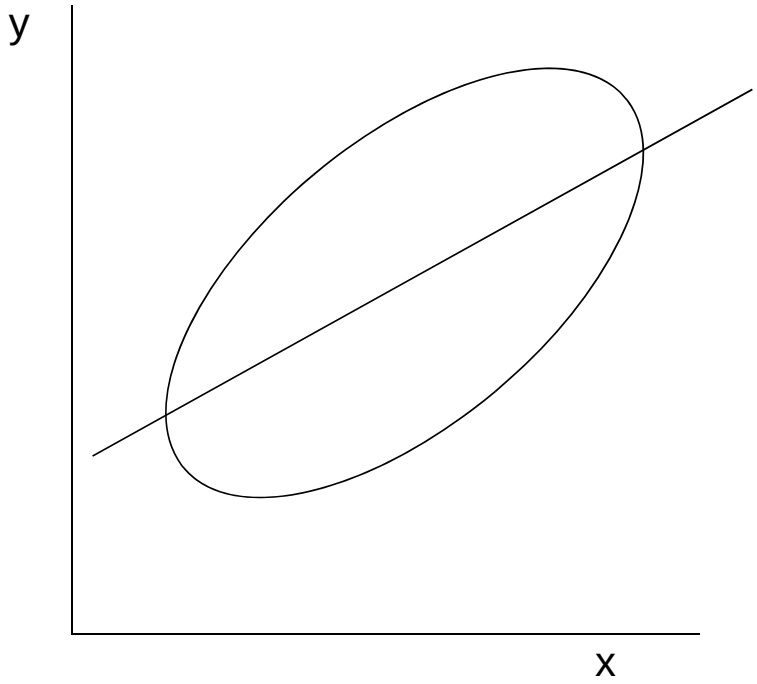
- Obvious
  - influences marginals
  - introduces variability into estimates
- Less obvious
  - Allows effective use of time and effort
  - Effect on multivariate techniques
    - Sampling of independent variable: greater precision in regression estimates
    - Sampling on dependent variable: bias

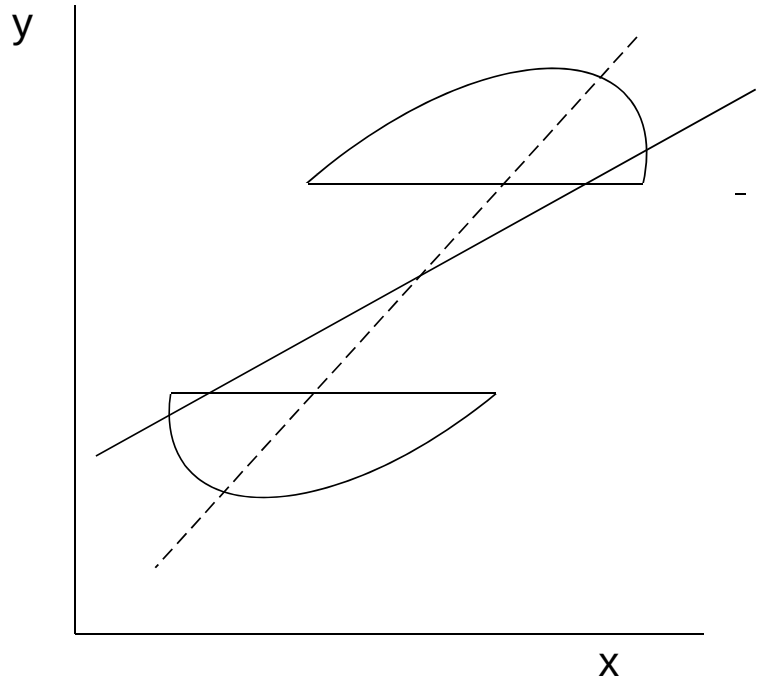
Sampling on Independent Variable





Sampling on Dependent Variable





Sampling

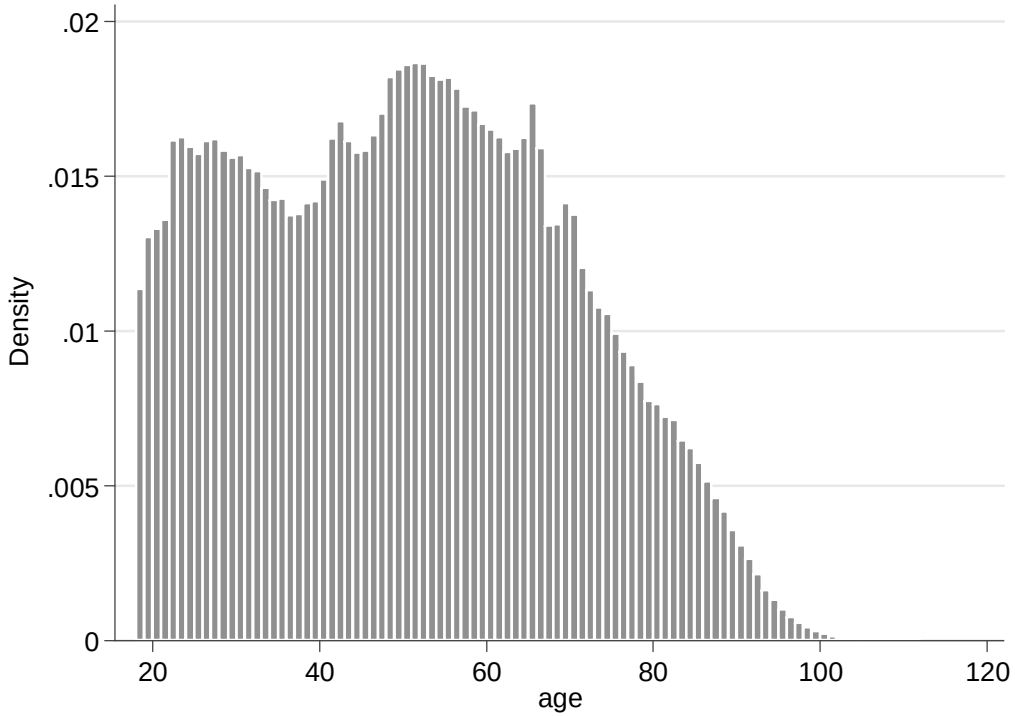
Consequences for Statistical Inference

Statistical Inference:
Learning About the Unknown From the Known

- Reasoning forward: distributions of sample means, when the population mean, s.d., and *n* are known.
- Reasoning backward: learning about the population mean when only the sample, s.d., and *n* are known

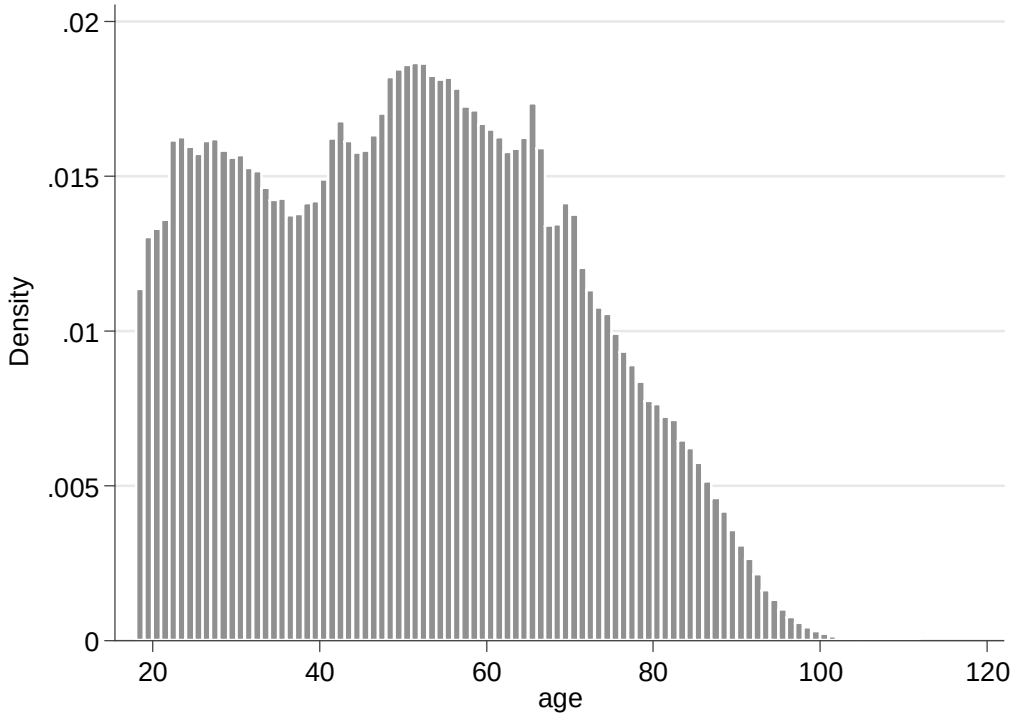
Reasoning Forward

Example: Age of Registered Voters in Florida




```
. summ age if age<=112&age>=18

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
         age |  12465018    50.22959    18.98939         18        112
```

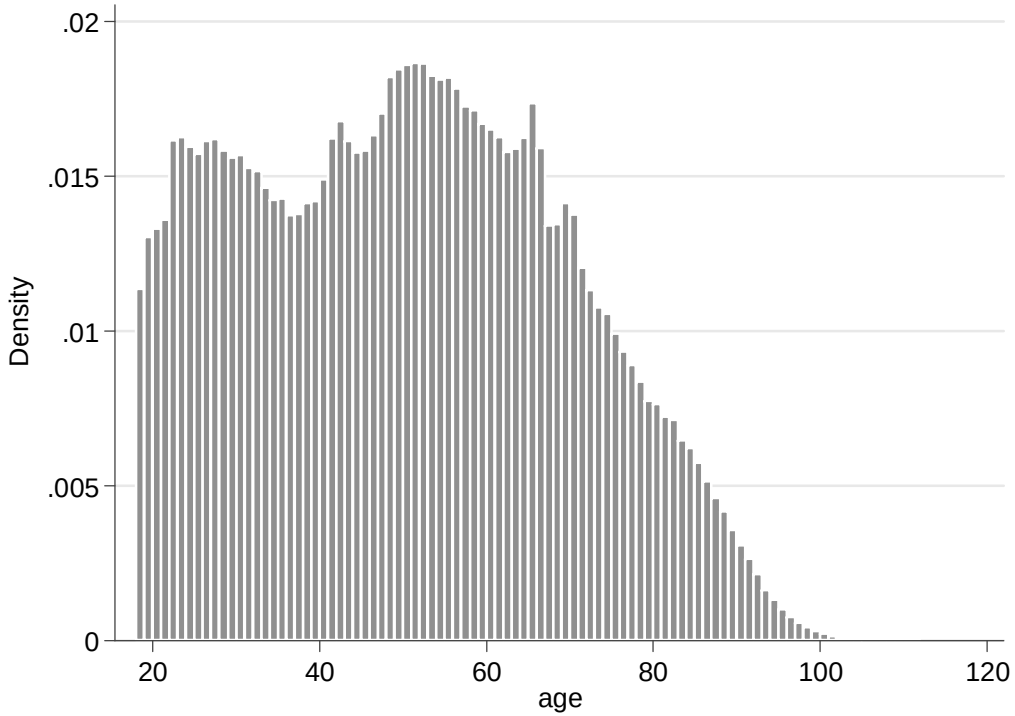
Example: Age of Registered Voters in Florida (Samples of 100 random observations)



Sample #	Sample mean	Running mean	Running s.d.	Abs. diff. from 50.23
1	50.36	50.36		0.13
2	47.63	49.00	0.97	1.23
3	50.32	49.44	0.70	0.79
4	53.42	50.43	0.71	0.20
5	50.84	50.51	0.69	0.28
6	48.57	50.19	0.62	0.04
7	47.3	49.78	0.57	0.45
8	48.33	49.60	0.55	0.63
9	49.95	49.64	0.52	0.59
10	50.66	49.74	0.49	0.49
11	51.7	49.92	0.47	0.31
12	52.7	50.15	0.45	0.08
13	50.41	50.17	0.44	0.06
14	50.99	50.23	0.43	0.00
15	50.51	50.25	0.42	0.02

```
. summ age if age<=112&age>=18

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
         age |  12465018    50.22959    18.98939         18        112
```

Example: Age of Registered Voters in Florida (Samples of 1000 random observations)



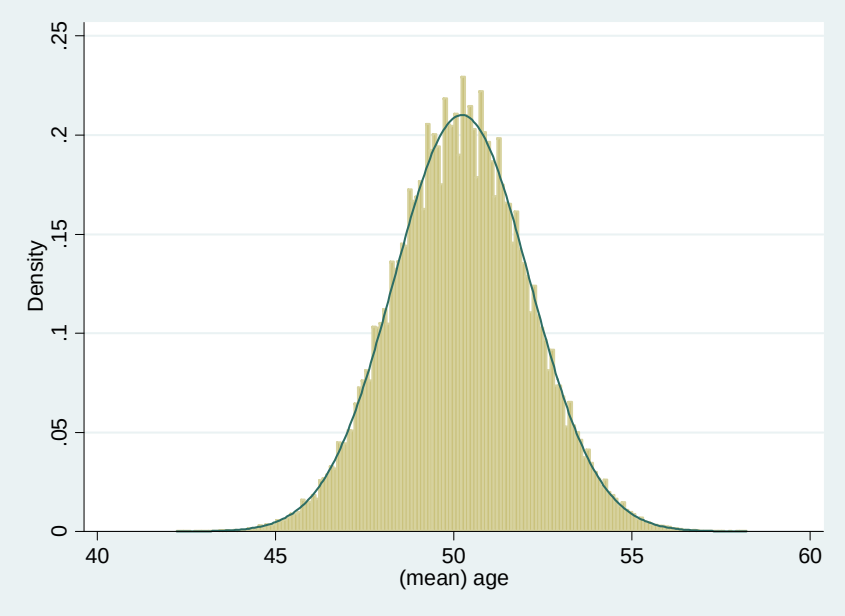
Sample #	Sample mean	Running mean	Running s.d.	Abs. diff. from 50.23
1	50.419	50.42		0.19
2	49.855	50.14	0.20	0.09
3	50.589	50.29	0.14	0.06
4	50.23	50.27	0.12	0.04
5	49.894	50.20	0.11	0.03
6	49.777	50.13	0.11	0.10
7	50.746	50.22	0.10	0.01
8	50.554	50.26	0.09	0.03
9	50.687	50.31	0.09	0.08
10	49.614	50.24	0.09	0.01
11	50.932	50.30	0.08	0.07
12	50.669	50.33	0.08	0.10
13	50.071	50.31	0.08	0.08
14	49.382	50.24	0.08	0.01
15	50.465	50.26	0.07	0.03

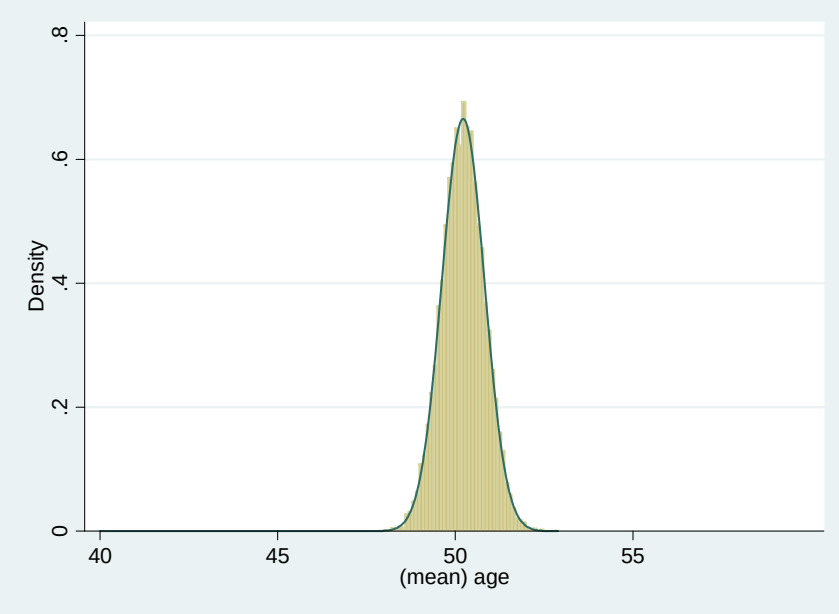
```
. summ age if age<=112&age>=18

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
         age |  12465018    50.22959    18.98939         18        112
```

Comparison of the 100 n samples and the 1000 n samples

Sample #	Sample mean	Running mean	Running s.d.	Abs. diff. from 50.23
1	50.36	50.36		0.13
2	47.63	49.00	0.97	1.23
3	50.32	49.44	0.70	0.79
4	53.42	50.43	0.71	0.20
5	50.84	50.51	0.69	0.28
6	48.57	50.19	0.62	0.04
7	47.3	49.78	0.57	0.45
8	48.33	49.60	0.55	0.63
9	49.95	49.64	0.52	0.59
10	50.66	49.74	0.49	0.49
11	51.7	49.92	0.47	0.31
12	52.7	50.15	0.45	0.08
13	50.41	50.17	0.44	0.06
14	50.99	50.23	0.43	0.00
15	50.51	50.25	0.42	0.02

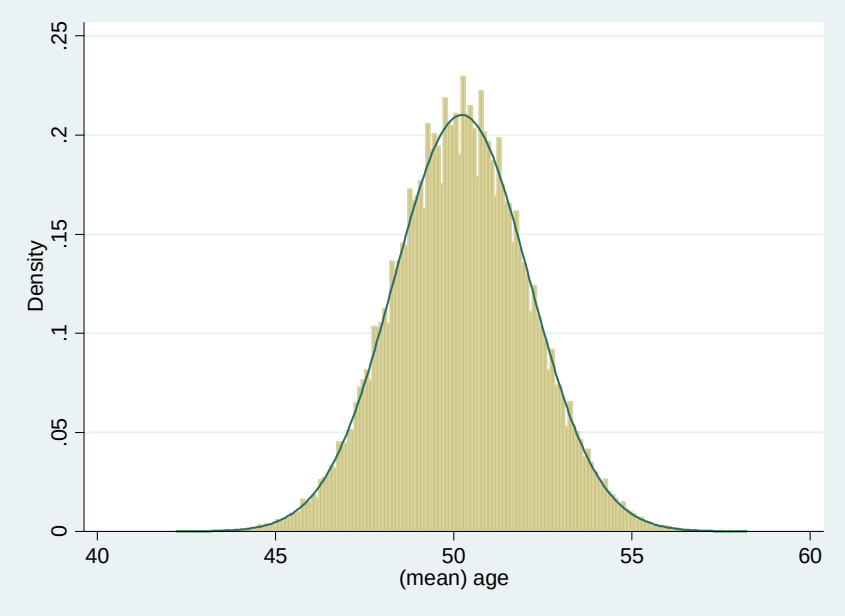
Sample #	Sample mean	Running mean	Running s.d.	Abs. diff. from 50.23
1	50.419	50.42		0.19
2	49.855	50.14	0.20	0.09
3	50.589	50.29	0.14	0.06
4	50.23	50.27	0.12	0.04
5	49.894	50.20	0.11	0.03
6	49.777	50.13	0.11	0.10
7	50.746	50.22	0.10	0.01
8	50.554	50.26	0.09	0.03
9	50.687	50.31	0.09	0.08
10	49.614	50.24	0.09	0.01
11	50.932	50.30	0.08	0.07
12	50.669	50.33	0.08	0.10
13	50.071	50.31	0.08	0.08
14	49.382	50.24	0.08	0.01
15	50.465	50.26	0.07	0.03

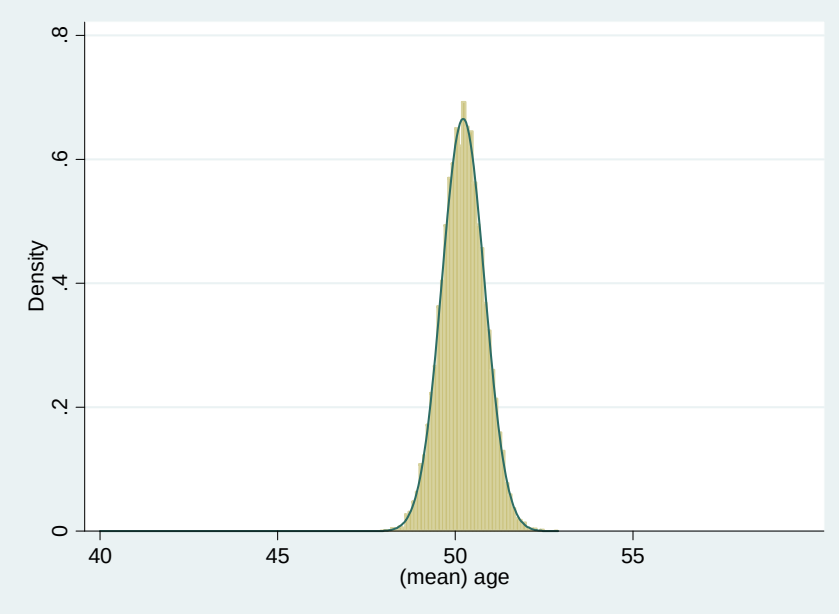





```
. tabstat age,statistics(count mean sd min max)

    variable |         N      mean        sd       min       max
-------------+--------------------------------------------------------
         age |    124650   50.2296  1.898515     42.22     58.14
----------------------------------------------------------------------
```

```
. tabstat age,statistics(count mean sd min max)

    variable |         N      mean        sd       min       max
-------------+--------------------------------------------------------
         age |     12465  50.22985  .5991014    47.944  52.84211
----------------------------------------------------------------------
```



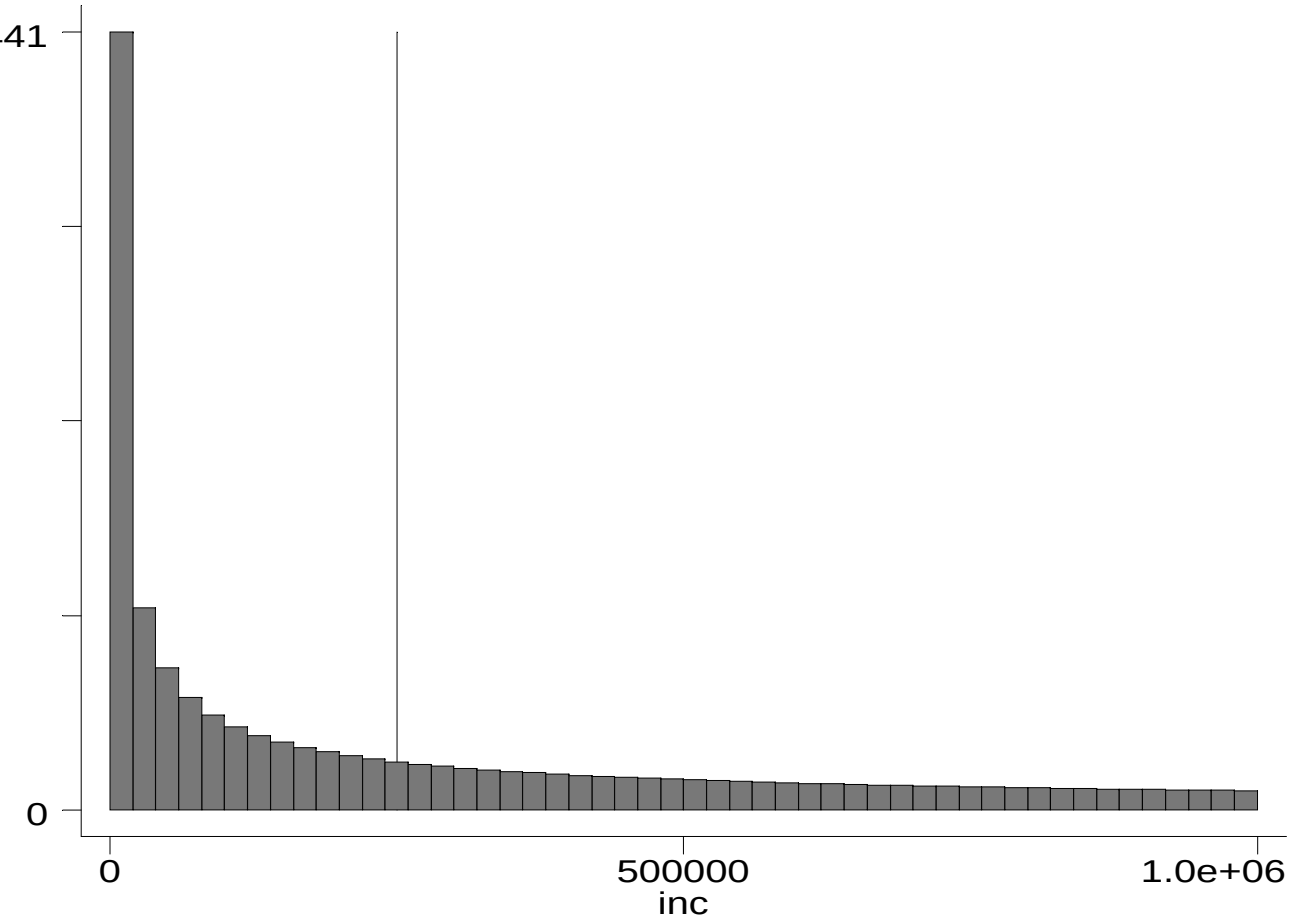



```
. tabstat age,statistics(count mean sd min max)

    variable |         N      mean        sd       min       max
-------------+--------------------------------------------------
         age |    124650   50.2296  1.898515     42.22     58.14
----------------------------------------------------------------
```

Standard error

```
. tabstat age,statistics(count mean sd min max)

    variable |         N      mean        sd       min       max
-------------+--------------------------------------------------
         age |     12465  50.22985  .5991014    47.944  52.84211
----------------------------------------------------------------
```

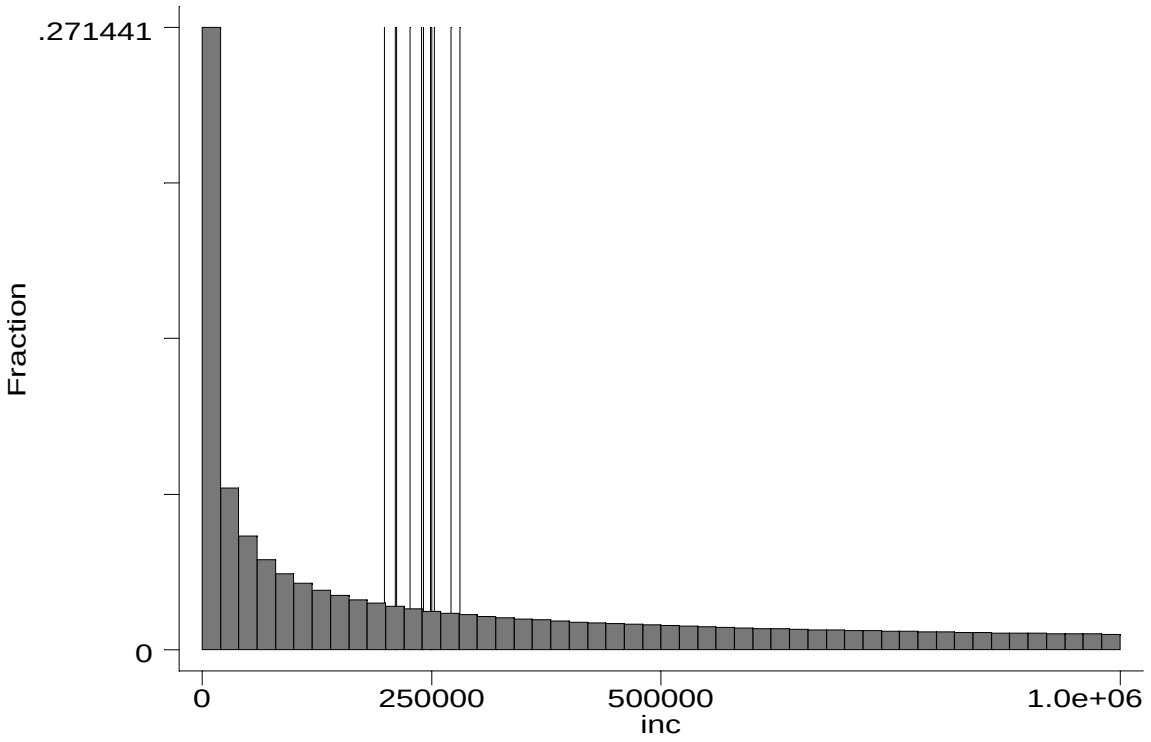
Exponential Distribution Example



Mean = 250,000
Median=125,000
s.d. = 283,474
Min = 0
Max = 1,000,000

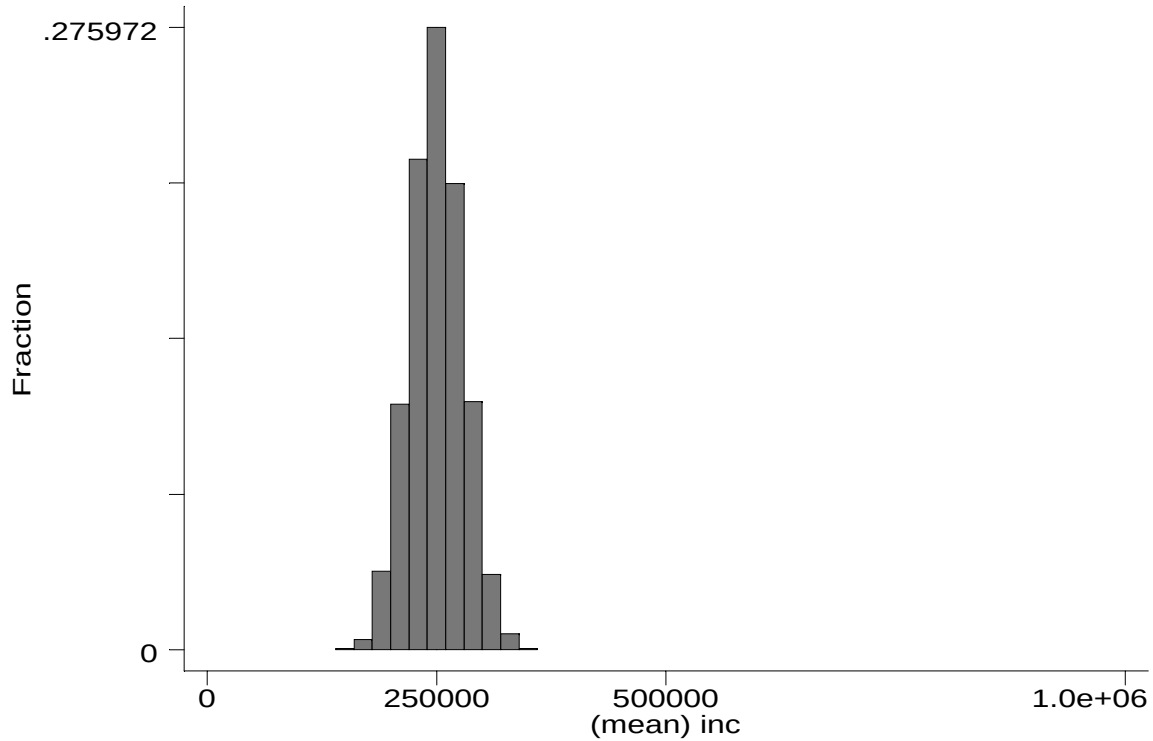
Consider 10 random samples, of *n* = 100 apiece

Sample	mean
1	253,396.9
2	198.789.6
3	271,074.2
4	238,928.7
5	280,657.3
6	241,369.8
7	249,036.7
8	226,422.7
9	210,593.4
10	212,137.3

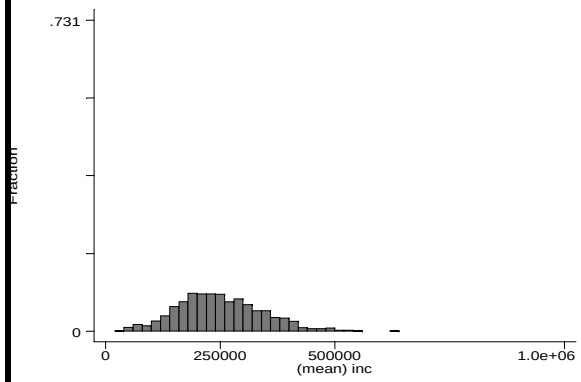
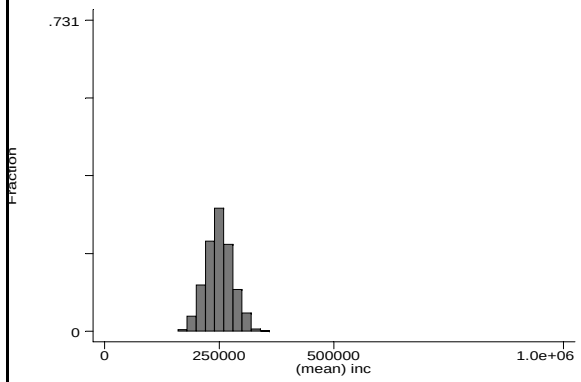
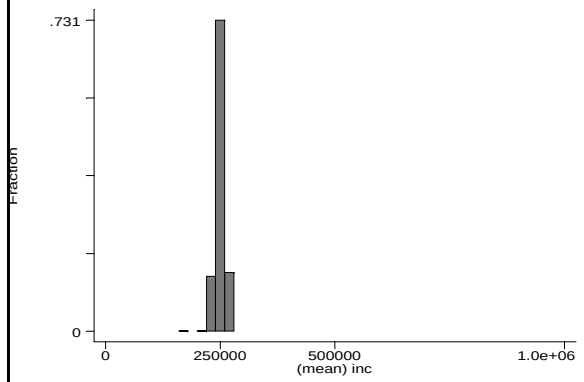


Consider 10,000 samples of *n* = 100

N = 10,000
Mean = 249,993
s.d. = 28,559
Skewness = 0.060
Kurtosis = 2.92



Consider 1,000 samples of various sizes

10	100	1000
		
Mean =250,105 s.d.= 90,891 Skew= 0.38 Kurt= 3.13	Mean = 250,498 s.d.= 28,297 Skew= 0.02 Kurt= 2.90	Mean = 249,938 s.d.= 9,376 Skew= -0.50 Kurt= 6.80

Reasoning Backward

When you know n, $\overline{X}$, and s,

but want to say something about μ

Reasoning Backward

When you know n, $\overline{X}$, and s,

but want to say something about μ

In other words, if you know the average age drawn from a sample (with n = 1,000) of the Florida voter registration file is 49.34, with a standard deviation of 21.81, what can we say about the likely *population* mean?

Central Limit Theorem

As the sample size *n* increases, the distribution of the mean $\bar{X}$ of a random sample taken from **practically any population** approaches a *normal* distribution, with mean μ and standard deviation $\sigma/\sqrt{n}$

Calculating Standard Errors

In general:

$$\text{std. err.} = \frac{s}{\sqrt{n}}$$

To answer the question…

In other words, if you know the average age drawn from a sample (with n = 1,000) of the Florida voter registration file is 49.34, with a standard deviation of 21.81, what can we say about the likely *population* mean?

To answer the question…

In other words, if you know the average age drawn from a sample (with n = 1,000) of the Florida voter registration file is 49.34, with a standard deviation of 21.81, what can we say about the likely *population* mean?

The most likely value of the population of the mean is 49.34. If we draw repeated samples of 1,000, we expect the standard deviation of those draws (i.e., the standard error) to be:

$$21.81 / \sqrt{1{,}000} = 0.69$$

Most important standard errors

Mean	$\frac{s}{\sqrt{n}}$
Proportion	$\sqrt{\frac{p(1-p)}{n}}$
Diff. of 2 means	$\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}$
Diff. of 2 proportions	$\sqrt{\frac{p_1(1-p_1)}{n_1}+\frac{p_2(1-p_2)}{n_2}}$
Diff of 2 means (paired data)	$\frac{s_d}{\sqrt{n}}$
Regression (slope) coeff.	$\frac{s.e.r.}{\sqrt{n-1}}\times\frac{1}{s_x}$

Standard error of a mean

$$\frac{s}{\sqrt{n}}$$

Example:
We have drawn a sample of 1,751 in-person Florida voters, in which they report waiting an average of 36.87 minutes to vote in 2012, with a standard deviation of 60.87 minutes. What is the standard error?

$$60.87 \Big/ \sqrt{1{,}751} = 1.45$$

Standard error of a proportion

$$\sqrt{\frac{p(1-p)}{n}}$$

Example:
We have drawn a sample of 2,578 Florida voters, in which 54% of them reported they voted for Obama instead of Romney. (Voters for other candidate have been excluded.) What is the standard error?

A: $\sqrt{.54(1-.54)/2578} = 0.0098 = 0.98\%$

Standard error of a difference of means

$$\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Example:

We have drawn a sample of 1,751 in-person Florida voters, in which they report waiting an average of 36.87 minutes to vote in 2012, with a standard deviation of 60.87 minutes. A sample of 1,104 Georgians reports an average waiting time of 16.02 with a standard deviation of 25.29 minutes. The difference of the average is 20.85 minutes. What is the standard error of this difference?

$$\sqrt{\frac{3704.70}{1{,}751} + \frac{639.83}{1{,}104}} = 1.64$$

Standard error of a difference of proportions

$$\sqrt{\frac{p_1(1-p_1)}{n_1}+\frac{p_2(1-p_2)}{n_2}}$$

Example:

We have drawn a sample of 2,578 Florida voters, in which 54% of the respondents said they voted for Obama. We have drawn another sample of 1,134 Georgia voters, in which 49% said they voted for Obama. The difference of the proportions is average is 5%. What is the standard error of this difference?

$$\sqrt{\frac{.54(1-.54)}{2{,}578}+\frac{.49(1-.49)}{1{,}134}}=0.018=1.8\%$$

Standard error of a regression coefficient

$$\frac{s.e.r.}{\sqrt{n-1}} \times \frac{1}{s_x}$$

Example:

We have drawn a sample of 36,906 voters nationwide. We have regressed a dummy variable equal to 1 if the respondent voted for Obama, 0 if for Romney. The independent variable is a 7-point party ID scale. All the relevant information is on the next slide. What is the standard error of the regression coefficient?

```
. reg obamavote dem7 [aw=V103]
(sum of wgt is   3.8204e+04)

      Source |       SS       df       MS              Number of obs =   36906
-------------+------------------------------           F(  1, 36904) =68743.59
       Model |  5976.53244     1  5976.53244           Prob > F      =  0.0000
    Residual |  3208.41472 36904  .086939484           R-squared     =  0.6507
-------------+------------------------------           Adj R-squared =  0.6507
       Total |  9184.94716 36905  .248880834           Root MSE      =  .29486

------------------------------------------------------------------------------
   obamavote |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        dem7 |   .1796006
       _cons |  -.2116884
------------------------------------------------------------------------------
```

$$\frac{s.e.r.}{\sqrt{n-1}} \times \frac{1}{s_x}$$

```
. tabstat dem7,statistics(sd)

    variable |        sd
-------------+----------
        dem7 |  2.239108
----------------------
```

$$\frac{0.29486}{\sqrt{36905}} \times \frac{1}{2.24} = 0.00069$$

```
. reg obamavote dem7 [aw=V103]
(sum of wgt is   3.8204e+04)

      Source |       SS       df       MS              Number of obs =   36906
-------------+------------------------------           F(  1, 36904) =68743.59
       Model |  5976.53244     1  5976.53244           Prob > F      =  0.0000
    Residual |  3208.41472 36904  .086939484           R-squared     =  0.6507
-------------+------------------------------           Adj R-squared =  0.6507
       Total |  9184.94716 36905  .248880834           Root MSE      =  .29486

------------------------------------------------------------------------------
   obamavote |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        dem7 |   .1796006    .000685
       _cons |  -.2116884   .0032303
------------------------------------------------------------------------------
```

$$\frac{s.e.r.}{\sqrt{n-1}} \times \frac{1}{s_x}$$

```
. tabstat dem7,statistics(sd)

    variable |        sd
-------------+----------
        dem7 |  2.239108
----------------------
```

$$\frac{0.29486}{\sqrt{36905}} \times \frac{1}{2.24} = 0.00069$$

```
. reg obamavote dem7 [aw=V103]
(sum of wgt is   3.8204e+04)

      Source |       SS       df       MS              Number of obs =   36906
-------------+------------------------------           F(  1, 36904) =68743.59
       Model |  5976.53244     1  5976.53244           Prob > F      =  0.0000
    Residual |  3208.41472 36904  .086939484           R-squared     =  0.6507
-------------+------------------------------           Adj R-squared =  0.6507
       Total |  9184.94716 36905  .248880834           Root MSE      =  .29486

------------------------------------------------------------------------------
   obamavote |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        dem7 |   .1796006    .000685
       _cons |  -.2116884   .0032303
------------------------------------------------------------------------------
```

variable	sd
dem7	2.239108

```
. reg obamavote dem7 [aw=V103] if dem7==1|dem7==7
(sum of wgt is   1.6532e+04)

      Source |       SS       df       MS              Number of obs =   17984
-------------+------------------------------           F(  1, 17982) =       .
       Model |  4195.86545     1  4195.86545           Prob > F      =  0.0000
    Residual |  287.126411 17982  .015967435           R-squared     =  0.9360
-------------+------------------------------           Adj R-squared =  0.9359
       Total |  4482.99186 17983  .249290545           Root MSE      =  .12636

------------------------------------------------------------------------------
   obamavote |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        dem7 |   .1612834   .0003146
       _cons |  -.1465254   .0016167
------------------------------------------------------------------------------
```

variable	sd
dem7	2.953895

Using Standard Errors, we can construct "confidence intervals"

- **Confidence interval (ci)**: an interval between two numbers, where there is a certain specified level of confidence that a population parameter lies

- ci = sample parameter ±

multiple * sample standard error

Using Standard Errors, we can construct "confidence intervals"

- **Confidence interval (ci)**: an interval between two numbers, where there is a certain specified level of confidence that a population parameter lies

- ci = sample parameter ±

multiple * sample standard error

Based on the normal curve

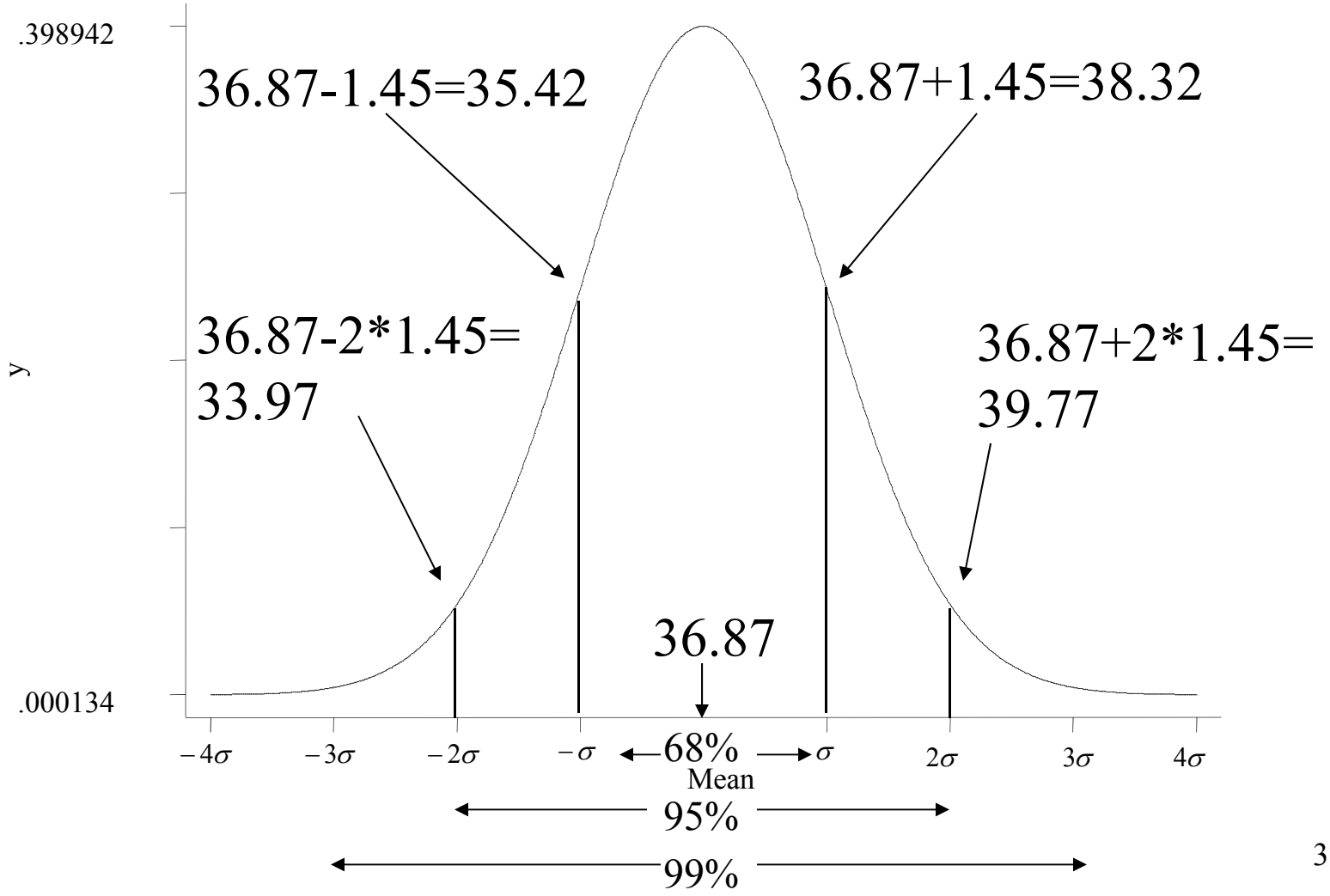
Constructing Confidence Intervals

Example:
We have drawn a sample of 1,751 in-person Florida voters, in which they report waiting an average of 36.87 minutes to vote in 2012, with a standard deviation of 60.87 minutes. What is the standard error?

$$60.87 \Big/ \sqrt{1{,}751} = 1.45$$

N = 1,751; avg. = 36.87; s.d. = 60.87; s.e. = s/√n = 1.45

The Picture



Confidence Intervals for Waiting Example

- 68% confidence interval = 36.87$\pm$ 1.45= [35.42 to 38.32]
- 95% confidence interval = 36.87$\pm$2*1.45 = [33.97 to 39.77]
- 99% confidence interval = 36.87$\pm$3*1.45 = [32.52 to 41.22]

What if someone (ahead of time) had said, "I think the average wait time to vote in Florida was 30 minutes."

- Note that 30 minutes is well out of the 99% confidence interval, [32.52 to 41.22]
- Q: How far away is the 45 minute estimate from the sample mean?
  - A: Do it in *z*-scores: (30-36.87)/1.45 = -4.73
- Consultation with an on-line z-distribution calculator reveals this is associated with the 99.999776% confidence interval.
- Alternatively, if the mean *were* really 30 min., the chance we would observe a sample mean this high is 100-99.999776 = 0.000224%

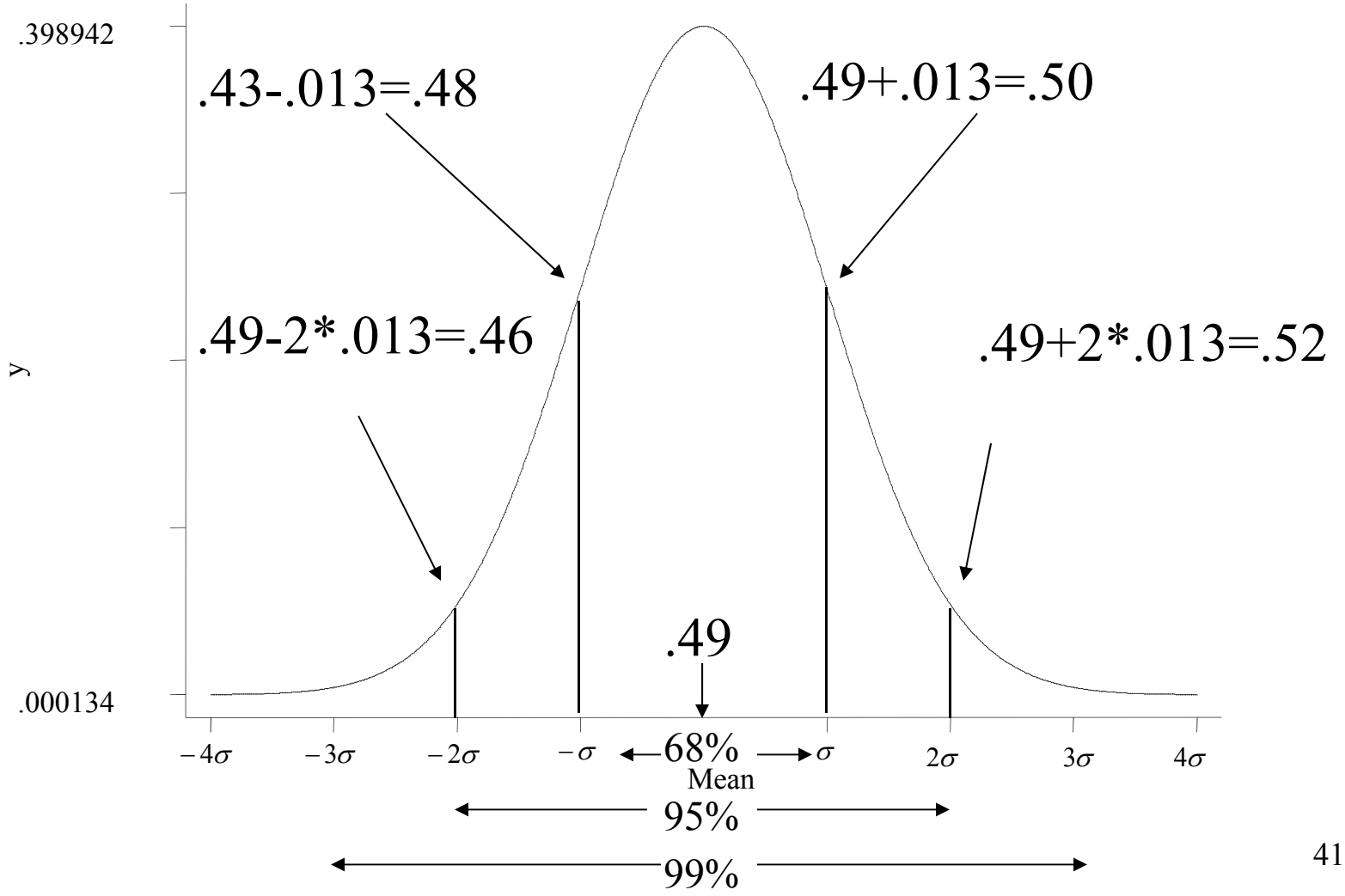
Constructing confidence intervals of proportions

- Let us say we drew a sample of 1500 adult nationwide and asked them if they approved of the way Barack Obama was handling his job as president. (April 6, 2013 Gallup Poll) Can we estimate the % of all American adults who approve?
- N = 1500
- p = .49
- s.e. = $\sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{.49(1-.49)}{1500}} = 0.013$

http://www.gallup.com/poll/113980/gallup-daily-obama-job-approval.aspx

N = 1,500; p. = .49; s.e. = $\sqrt{p(1-p)/n}$ = .013

The Picture



Confidence Intervals for Obama approval example

- 68% confidence interval = .49±.013 =

[.48 to .50]

- 95% confidence interval = .49±2*.013 =

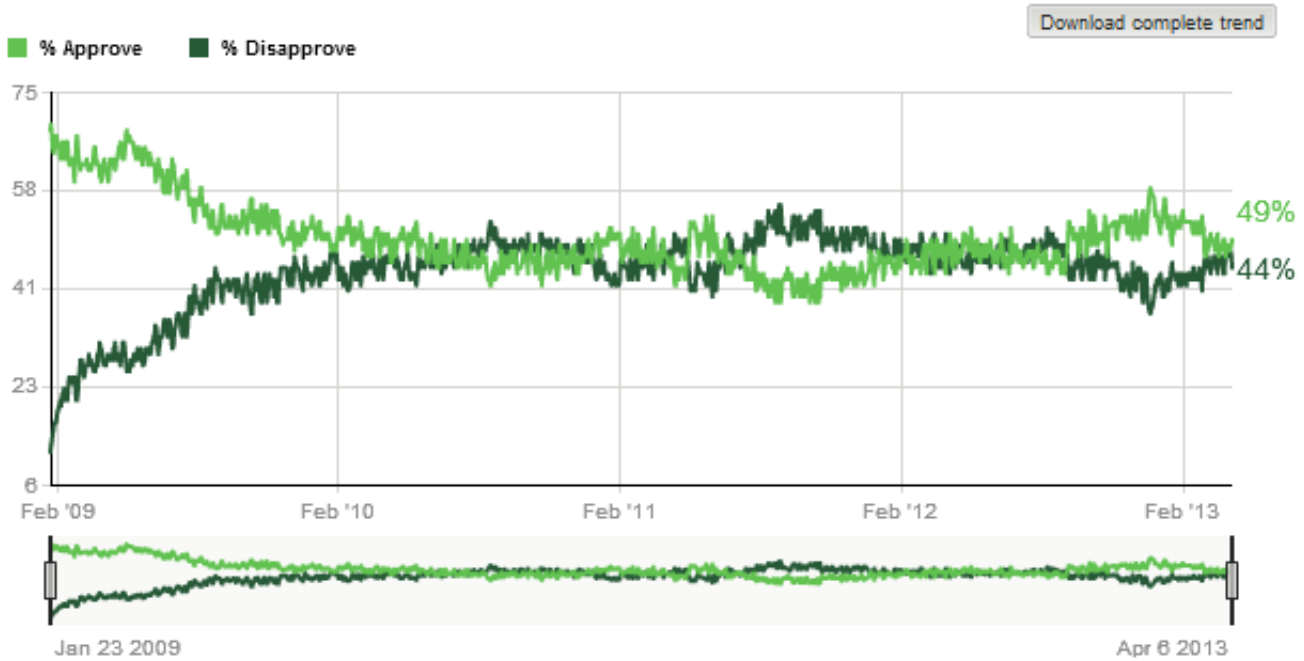
[.46 to .52]

- 99% confidence interval = .49±3*.013 =

[.45 to .53]

Gallup Daily: Obama Job Approval

Each result is based on a three-day rolling average



Gallup tracks daily the percentage of Americans who approve or disapprove of the job Barack Obama is doing as president. Daily results are based on telephone interviews with approximately 1,500 national adults; Margin of error is ±3 percentage points.

Really 2.6 percentage points

GALLUP® Politics

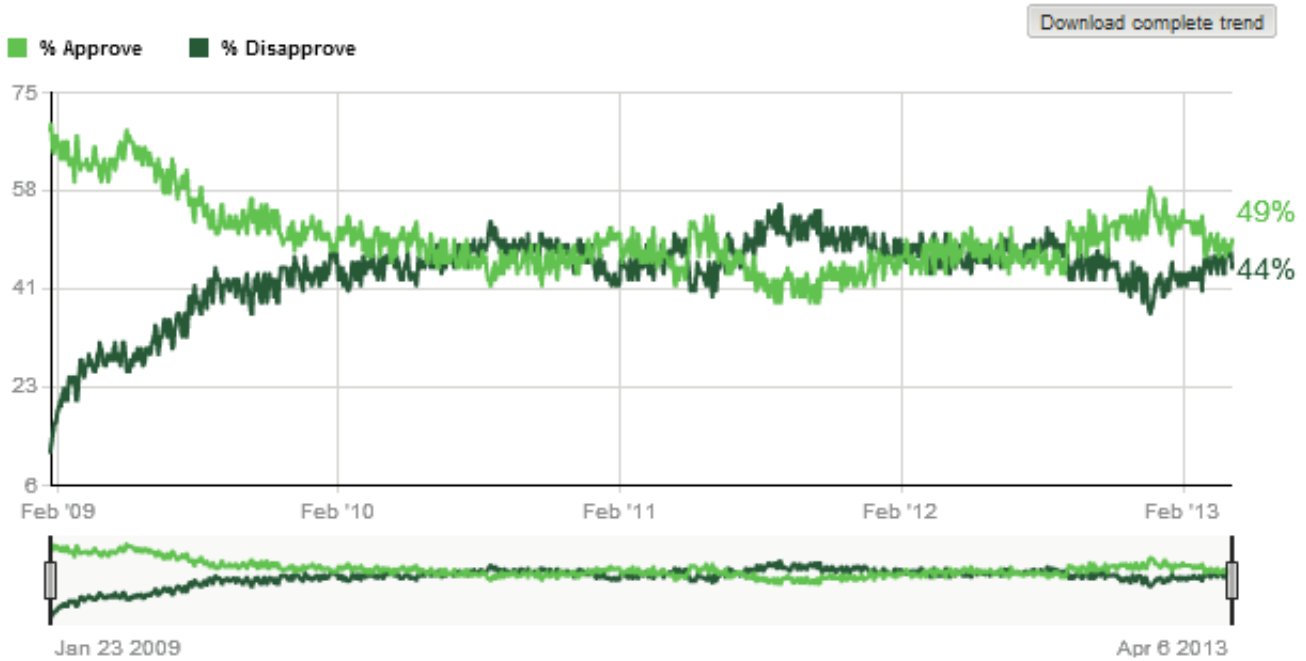
HOME | POLITICS | ECONOMY | WELLBEING | WORLD

HOT TOPICS: North Korea | Same-Sex Relations | Israeli-Palestinian Conflict | Catholics | Sequestration | State c

SHARE

Gallup Daily: Obama Job Approval

Each result is based on a three-day rolling average



Gallup tracks daily the percentage of Americans who approve or disapprove of the job Barack Obama is doing as president. Daily results are based on telephone interviews with approximately 1,500 national adults; Margin of error is ±3 percentage points.

What if someone (ahead of time) had said, "I think Americans are equally divided in how they think about Obama."

- Note that 50% is well within the 95% confidence interval, [46% to 52%]
- Q: How far away is the 50% estimate from the sample proportion?
  - A: Do it in *z*-scores: (.49-.50)/.013 = -0.77
- This is associated with the 56% confidence interval.

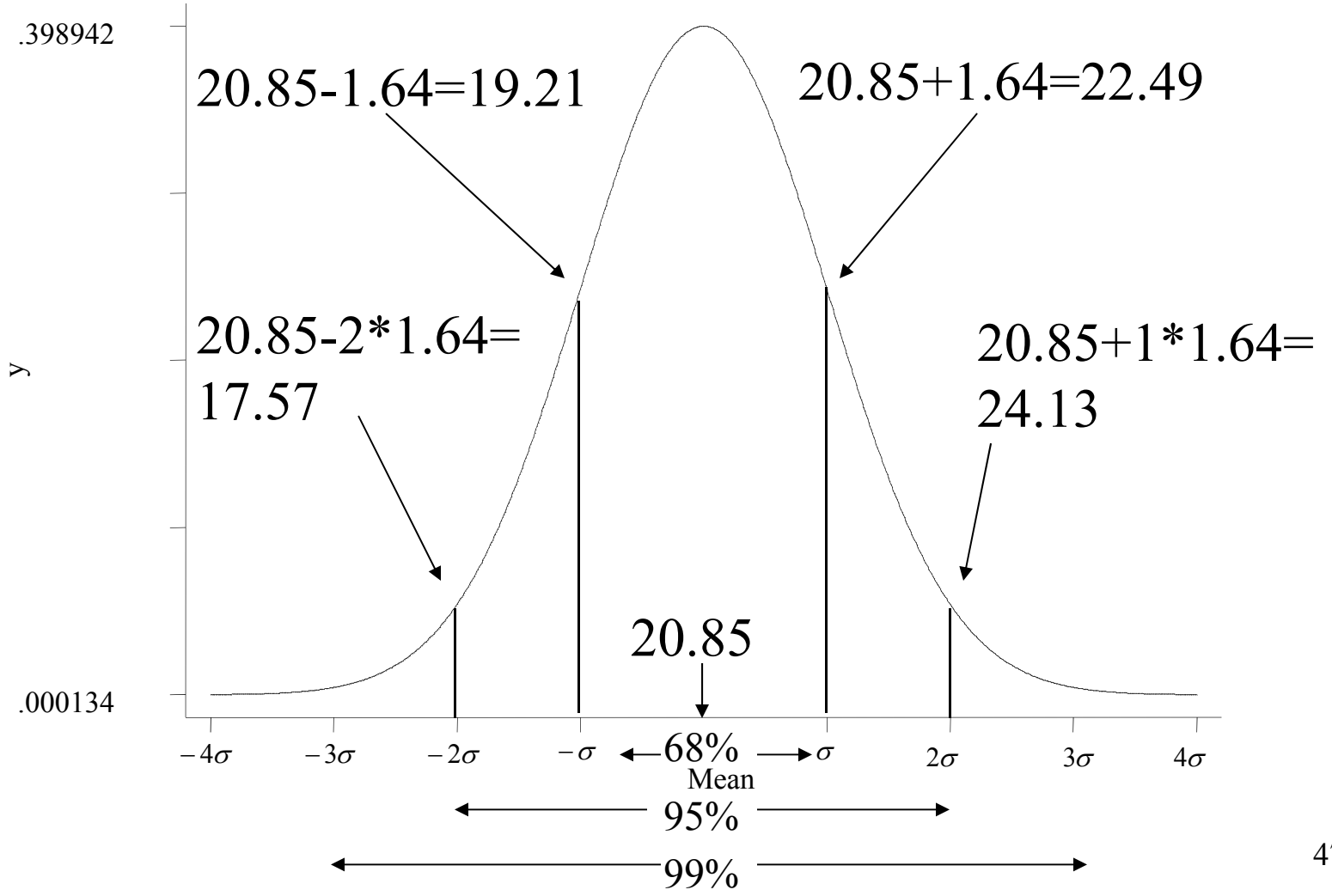
Constructing confidence intervals of differences of means

- Let us say we have drawn a sample of 1,751 in-person Florida voters, in which they report waiting an average of 36.87 minutes to vote in 2012, with a standard deviation of 60.87. A sample of 1,104 Georgians reports an average wait time of 16.02 with a standard deviation of 25.29 minutes. The difference of the average is 20.85 minutes.

- N = 1,751 for FL and 1,104 for GA
- Average = 36.87 (FL); 16.02 (GA); diff = 20.85
- s.d. = 60.87 (FL); 25.29 (GA)
- s.e. = $$\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = \sqrt{\frac{3704.70}{1{,}751} + \frac{639.83}{1{,}104}} = 1.64$$

N = 1,751 (FL); 1,104 (GA); diff = 20.85; s.e. = 1.64

The Picture



Confidence Intervals for difference of waiting times example

- 68% confidence interval = 20.85±1.64 =

[19.21 to 22.49]

- 95% confidence interval = 20.85±2*1.64 = [17.57 to 24.13]
- 99% confidence interval =20.85±3*1.64 = [15.93 to 25.77]

What if someone (ahead of time) had said, “People waited just as long on average in Georgia as in Florida”

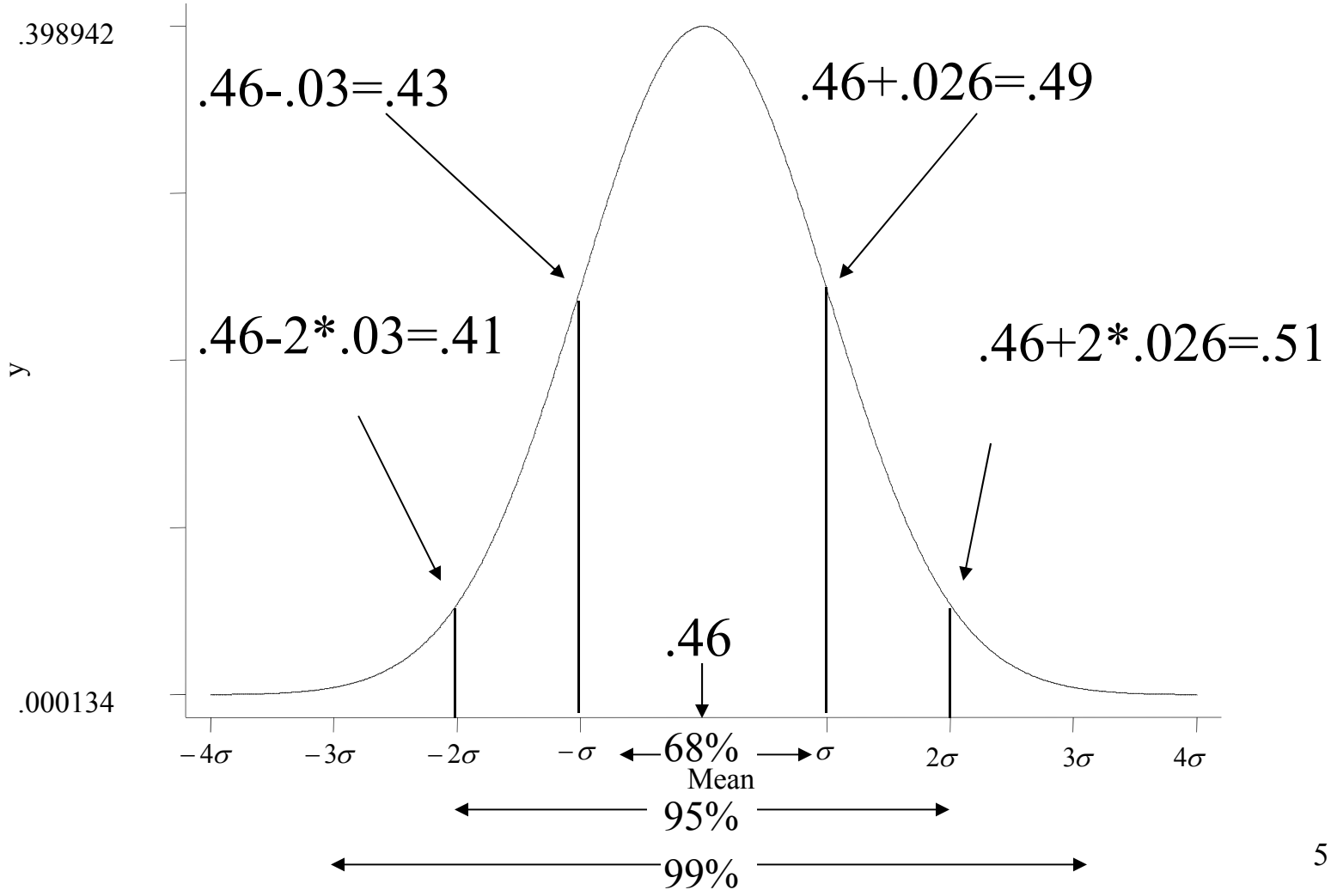
- Note that 0 is well out of the 95% confidence interval, [17.75 to 24.13]
- Q: How far away is the 0 estimate from the sample difference?
  - A: Do it in *z*-scores: (20.85)/1.64 = 12.7

Constructing confidence intervals of difference of proportions

- Let us say we drew a sample of 1,500 adults and asked them if they approved of the way Barack Obama was handling his job as president. We focus on the 1000 who are either independents or Democrats. Can we estimate whether independents and Democrats view Obama differently?
- N = 600 Ind.; 400 Dem.
- p = .42 (Ind.); .88 (Dem.); diff = .46
- s.e. = $$\sqrt{\frac{p_1(1-p_1)}{n_1}+\frac{p_2(1-p_2)}{n_2}}=\sqrt{\frac{.42(1-.42)}{600}+\frac{.88(1-.88)}{400}}=.026$$

diff. p. = .46; s.e. = .026

The Picture



Confidence Intervals for Obama Ind/Dem approval example

- 68% confidence interval = .46±.026 =

[.43 to .49]

- 95% confidence interval = .46±2*.026 =

[.41 to .51]

- 99% confidence interval = .46±3*.026 =

[.38 to .54]

What if someone (ahead of time) had said, “I think Democrats and Independents are equally unsupportive of Obama”?

- Note that 0% is well out of the 95% confidence interval, [41% to 51%]
- Q: How far away is the 0% estimate from the sample proportion?
  - A: Do it in *z*-scores: (.41-0)/.026 = 15.77

Constructing confidence intervals of regression coefficients

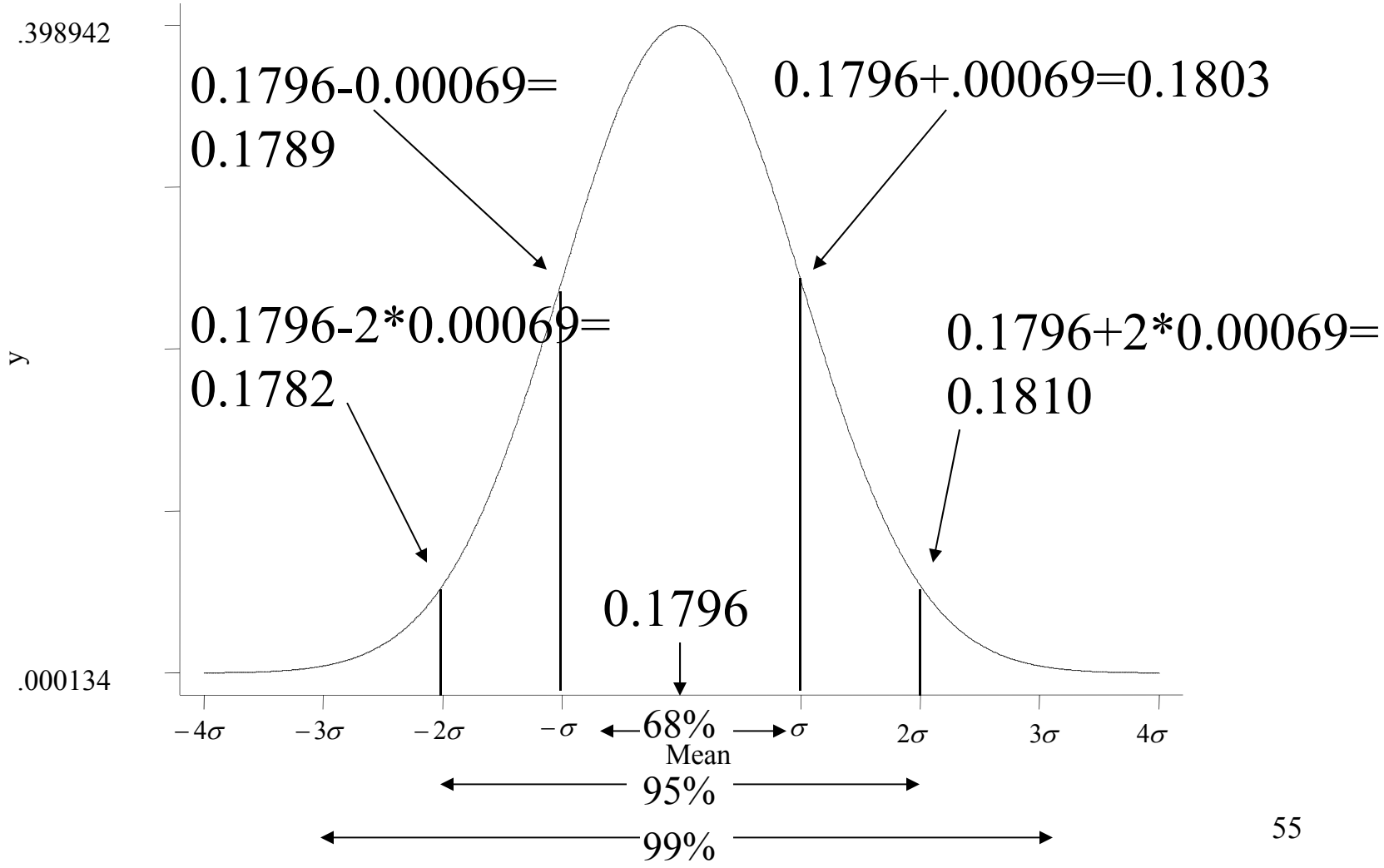
- Let's look at the relationship between party identification and the probability of voting for Obama.

Slope = 0.18
N = 36,906
s.e.r. = .29
$s_x = 2.24$
$s.e._{slope} = 0.00069$

$$\frac{s.e.r.}{\sqrt{n-1}} \times \frac{1}{s_x} = \frac{0.29}{\sqrt{36{,}905}} \times \frac{1}{2.24} = 0.00069$$

N = 36,908; slope=0.1796; s.e. = 0.00069

The Picture



Confidence Intervals for regression example

- 68% confidence interval = 0.1796± 0.00069=

[0.1789 to 0.1803]

- 95% confidence interval = 0.1796± 2*0.00069= [0.1782 to 0.1810]
- 99% confidence interval = 0.1796± 3*0.00069= [0.1775 to 0.1817]

What if someone (ahead of time) had said, “There is no relationship between a voter’s ideology and the probability of voting for Obama for President”?

- Note that 0 is well out of the 95% confidence interval, [0.1782 to 0.1810]
- Q: How far away is the 0 estimate from the sample proportion?
  - A: Do it in *z*-scores: (0.1796-0)/0.00069 = 260.29

The Stata output

```
. reg obamavote dem7 [aw=V103]
(sum of wgt is   3.8204e+04)

      Source |       SS       df       MS              Number of obs =   36906
-------------+------------------------------           F(  1, 36904) =68743.59
       Model |  5976.53244     1  5976.53244           Prob > F      =  0.0000
    Residual |  3208.41472 36904  .086939484           R-squared     =  0.6507
-------------+------------------------------           Adj R-squared =  0.6507
       Total |  9184.94716 36905  .248880834           Root MSE      =  .29486

------------------------------------------------------------------------------
   obamavote |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        dem7 |   .1796006    .000685   262.19   0.000      .178258    .1809432
       _cons |  -.2116884   .0032303   -65.53   0.000    -.2180198   -.2053569
------------------------------------------------------------------------------
```