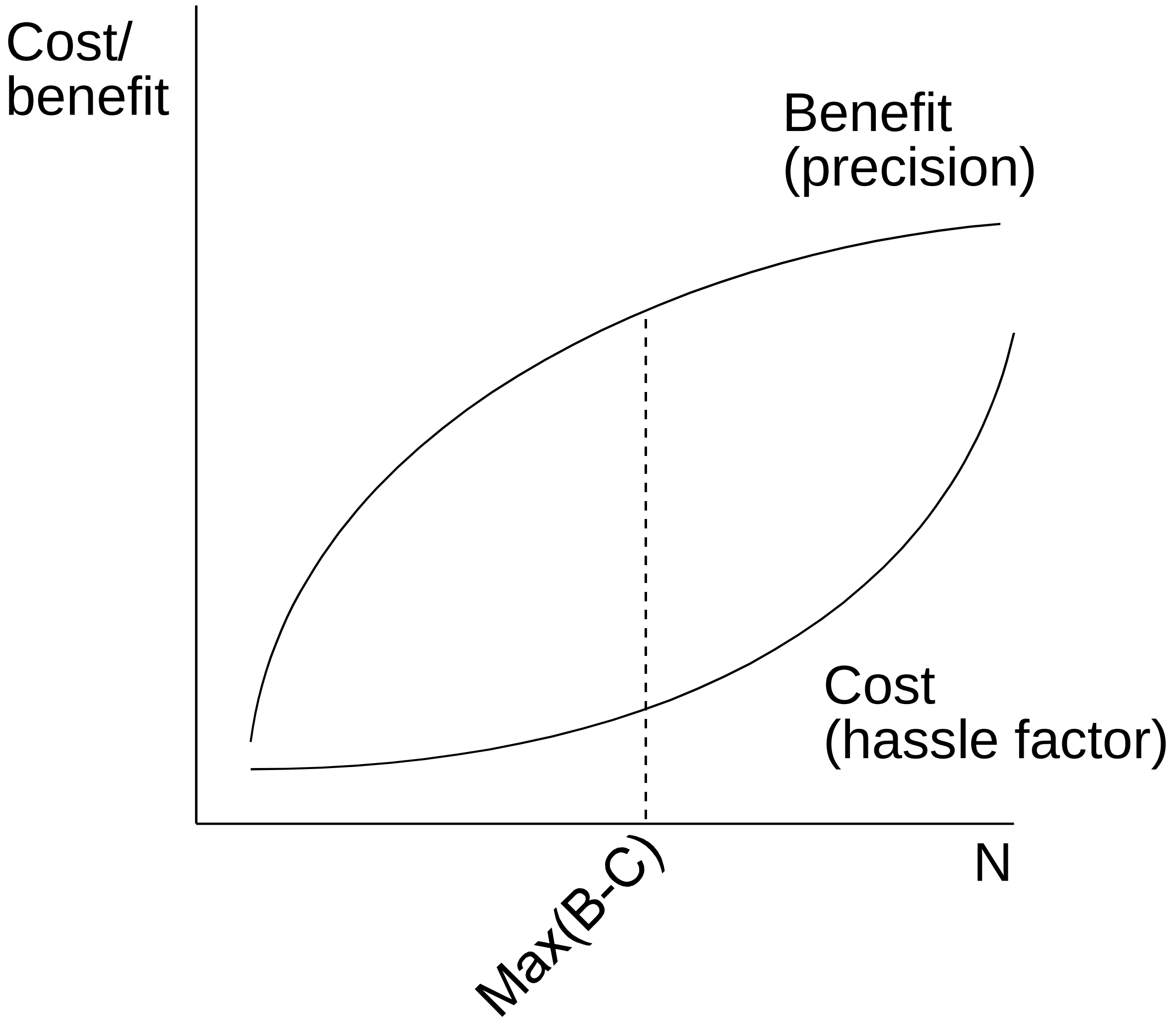
Sampling and Inference

The Quality of Data and Measures

2015

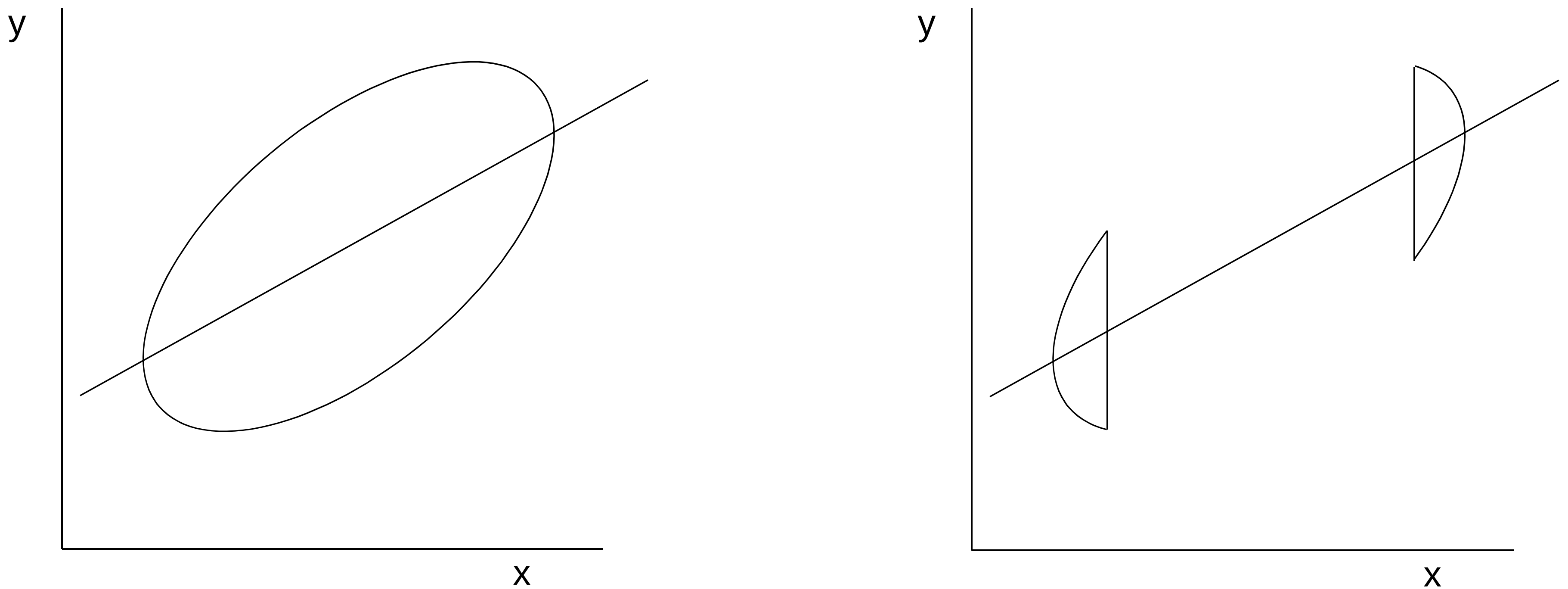
Why do we sample?



Effects of samples

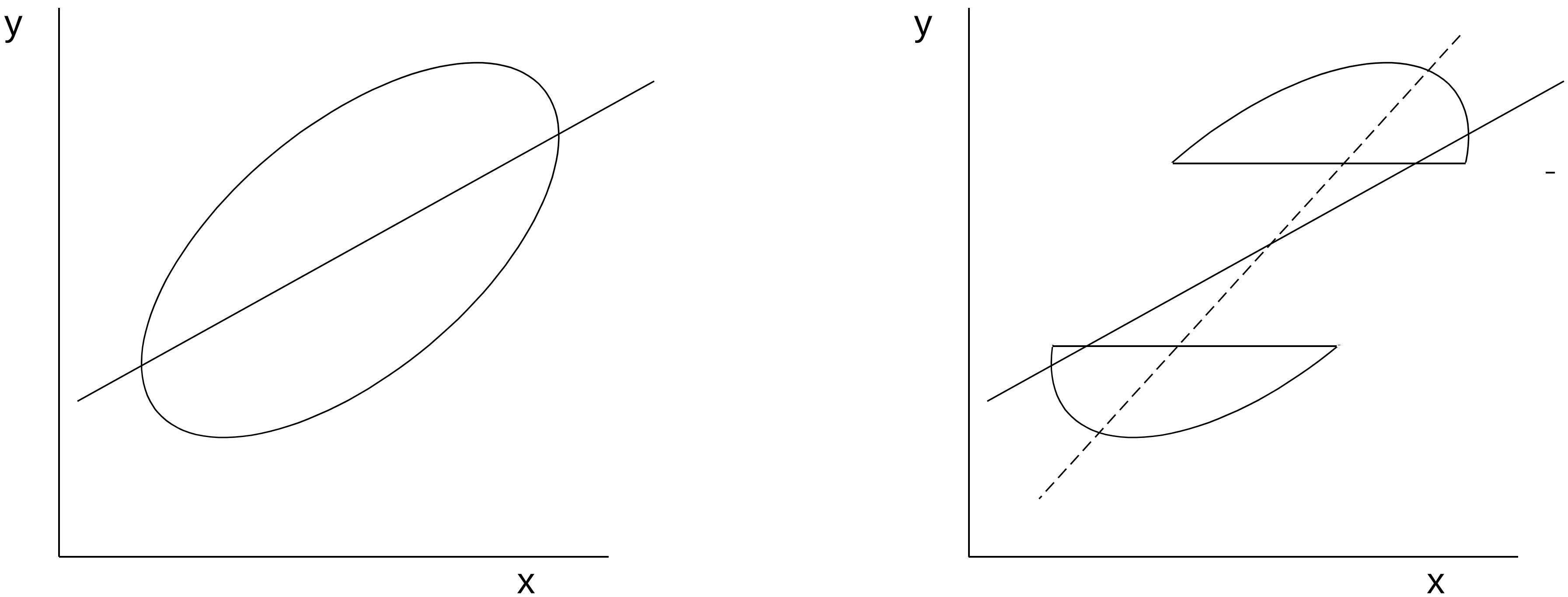
- Obvious
  - influences marginals
  - introduces variability into estimates
- Less obvious
  - Allows effective use of time and effort
  - Effect on multivariate techniques
    - Sampling of independent variable: greater precision in regression estimates
    - Sampling on dependent variable: bias

Sampling on Independent Variable



Because an assumption of OLS is that $E(y) = \beta_0 + \beta_1 x$ for all values of x.

Sampling on Dependent Variable



Because an assumption of OLS is that cov(x,ε) = 0

Sampling

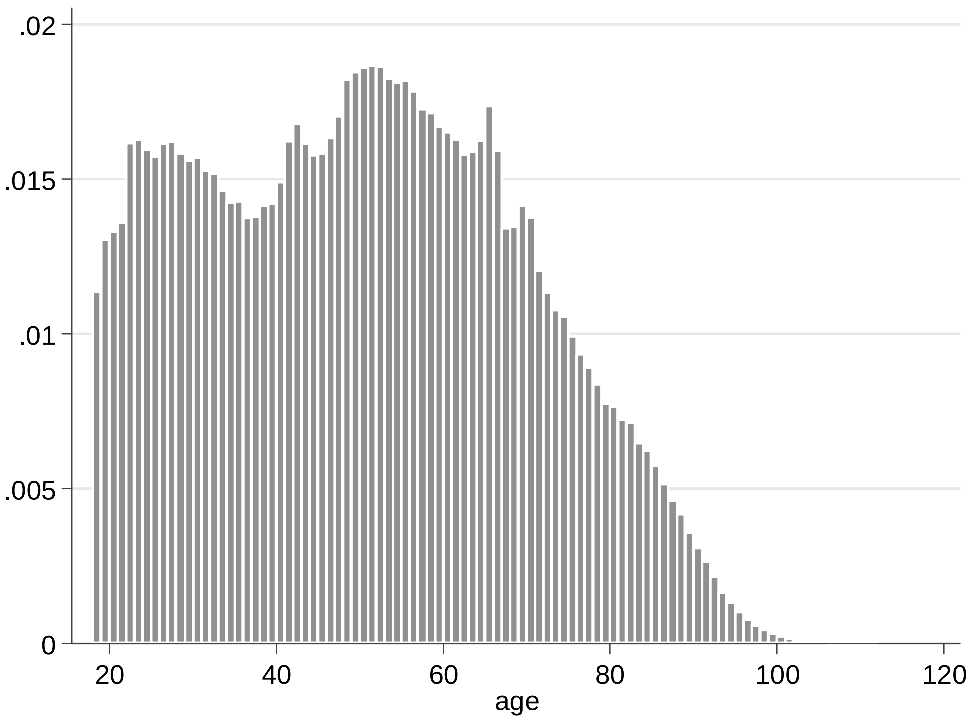
Consequences for Statistical Inference

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

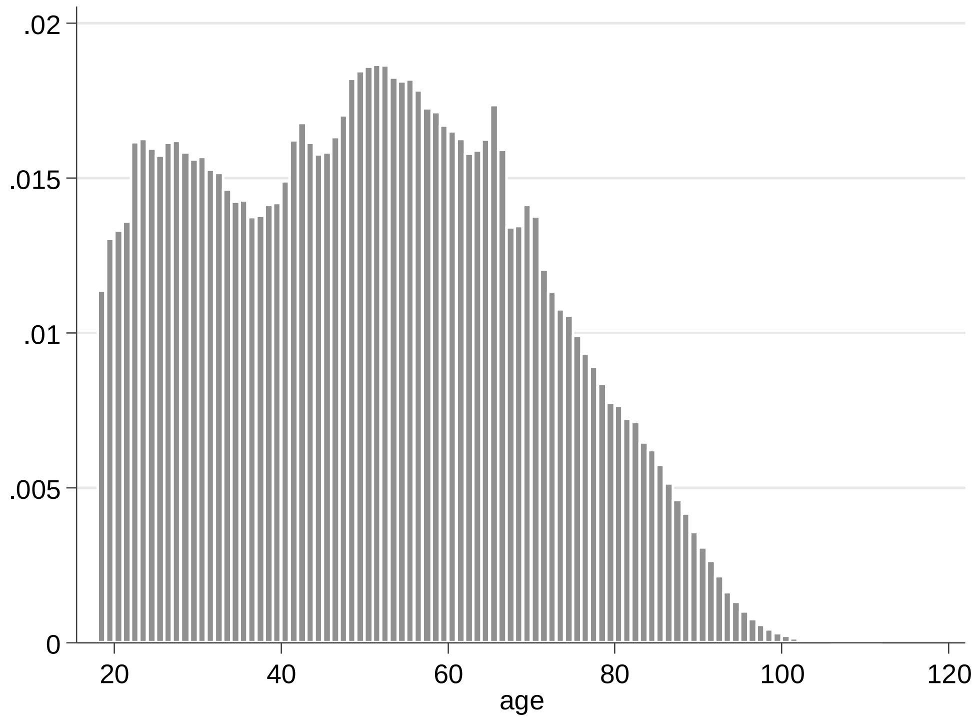
Reasoning Forward

Example: Age of Registered Voters in Florida




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

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

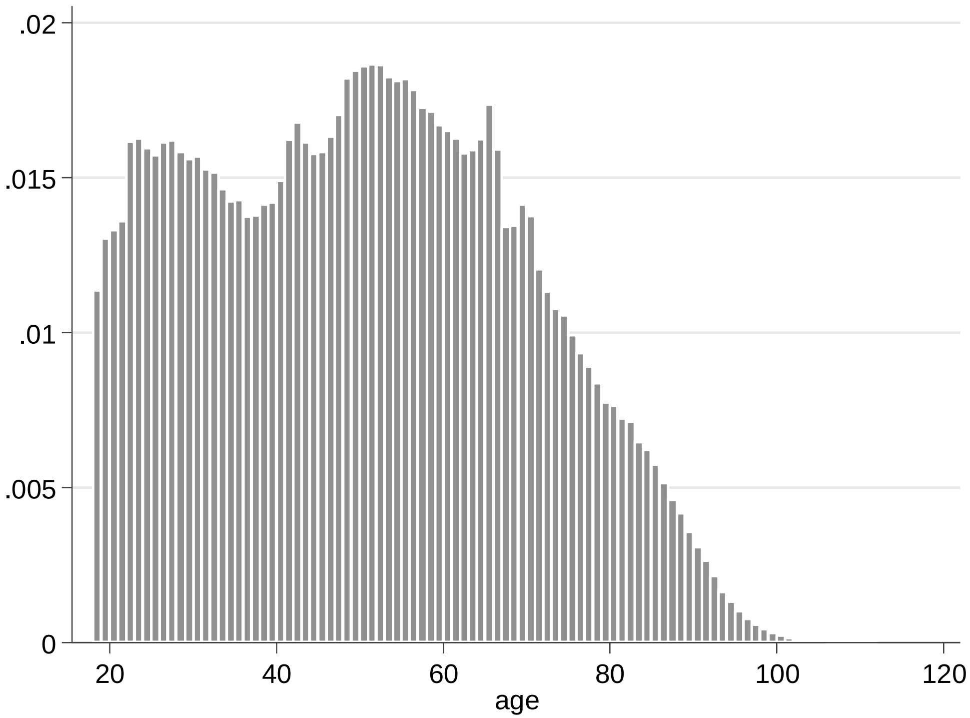
Example: Age of Registered Voters in Florida (Samples of 100 random observations)



sampleno	mean	running mean	abs. diff from 50.23	running s.d.
1	49.46	49.46	0.77	
2	50.26	49.86	0.37	0.57
3	47.50	49.07	1.16	1.42
4	50.60	49.46	0.77	1.39
5	53.13	50.19	0.04	2.04
6	50.90	50.31	0.08	1.84
7	48.72	50.08	0.15	1.79
8	46.89	49.68	0.55	2.00
9	48.52	49.55	0.68	1.91
10	49.87	49.59	0.64	1.81
11	51.01	49.71	0.52	1.77
12	51.38	49.85	0.38	1.75
13	52.70	50.07	0.16	1.85
14	50.59	50.11	0.12	1.79
15	51.21	50.18	0.05	1.74
16	50.27	50.19	0.04	1.69
17	53.14	50.36	0.13	1.78
18	48.85	50.28	0.05	1.77
19	52.53	50.40	0.17	1.79
20	47.60	50.26	0.03	1.85

```
. 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)



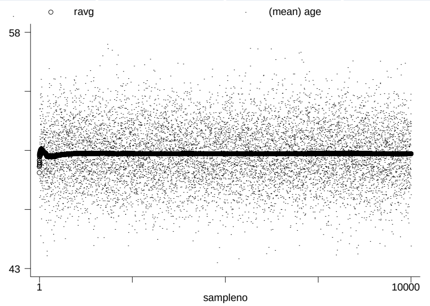
sampleno	mean	running mean	abs. diff from 50.23	running s.d.
1	49.59	49.59	0.64	
2	50.93	50.26	0.03	0.95
3	50.98	50.50	0.27	0.79
4	50.57	50.52	0.29	0.65
5	50.71	50.55	0.33	0.57
6	50.49	50.54	0.31	0.51
7	50.29	50.51	0.28	0.47
8	49.83	50.42	0.19	0.50
9	50.81	50.46	0.24	0.49
10	49.20	50.34	0.11	0.61
11	50.76	50.38	0.15	0.59
12	49.97	50.34	0.11	0.57
13	50.50	50.35	0.13	0.55
14	49.96	50.33	0.10	0.54
15	50.25	50.32	0.09	0.52
16	50.12	50.31	0.08	0.51
17	49.79	50.28	0.05	0.51
18	49.37	50.23	0.00	0.54
19	49.96	50.21	0.02	0.52
20	48.85	50.15	0.08	0.59

```
. 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

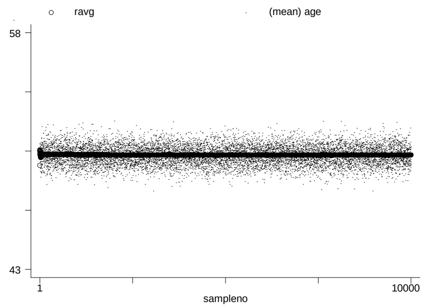
100

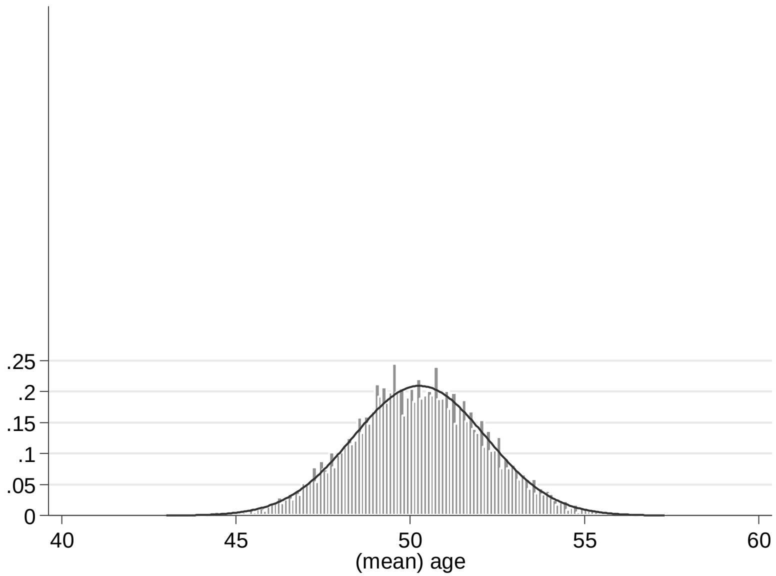
sampleno	mean	running mean	abs. diff from 50.23	running s.d.
1	49.46	49.46	0.77	
2	50.26	49.86	0.37	0.57
3	47.50	49.07	1.16	1.42
4	50.60	49.46	0.77	1.39
5	53.13	50.19	0.04	2.04
6	50.90	50.31	0.08	1.84
7	48.72	50.08	0.15	1.79
8	46.89	49.68	0.55	2.00
9	48.52	49.55	0.68	1.91
10	49.87	49.59	0.64	1.81
11	51.01	49.71	0.52	1.77
12	51.38	49.85	0.38	1.75
13	52.70	50.07	0.16	1.85
14	50.59	50.11	0.12	1.79
15	51.21	50.18	0.05	1.74
16	50.27	50.19	0.04	1.69
17	53.14	50.36	0.13	1.78
18	48.85	50.28	0.05	1.77
19	52.53	50.40	0.17	1.79
20	47.60	50.26	0.03	1.85

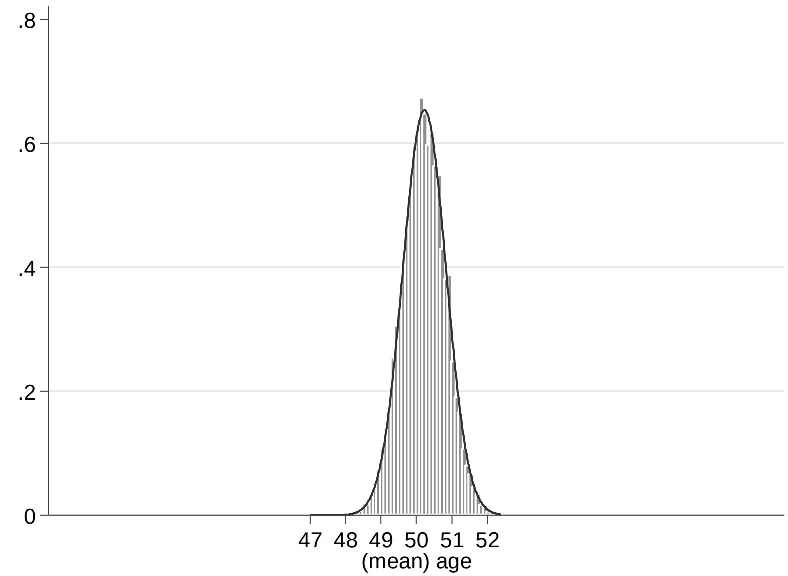


1,000

sampleno	mean	running mean	abs. diff from 50.23	running s.d.
1	49.59	49.59	0.64	
2	50.93	50.26	0.03	0.95
3	50.98	50.50	0.27	0.79
4	50.57	50.52	0.29	0.65
5	50.71	50.55	0.33	0.57
6	50.49	50.54	0.31	0.51
7	50.29	50.51	0.28	0.47
8	49.83	50.42	0.19	0.50
9	50.81	50.46	0.24	0.49
10	49.20	50.34	0.11	0.61
11	50.76	50.38	0.15	0.59
12	49.97	50.34	0.11	0.57
13	50.50	50.35	0.13	0.55
14	49.96	50.33	0.10	0.54
15	50.25	50.32	0.09	0.52
16	50.12	50.31	0.08	0.51
17	49.79	50.28	0.05	0.51
18	49.37	50.23	0.00	0.54
19	49.96	50.21	0.02	0.52
20	48.85	50.15	0.08	0.59

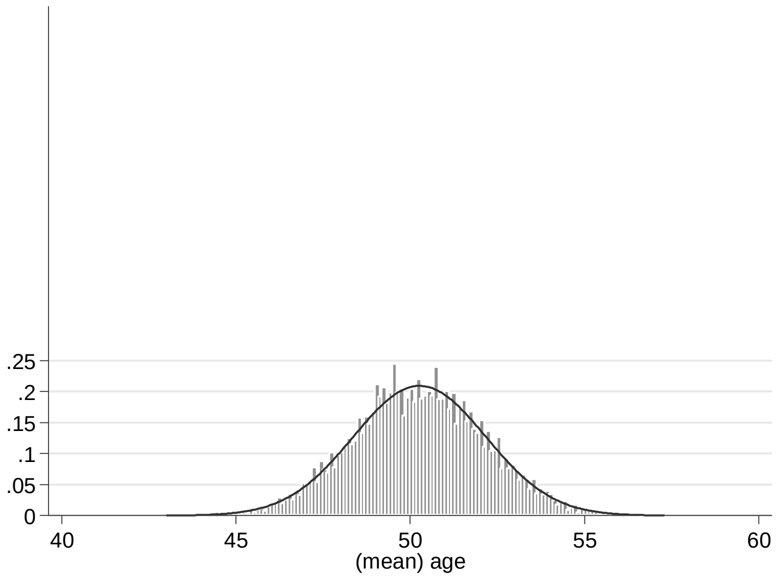


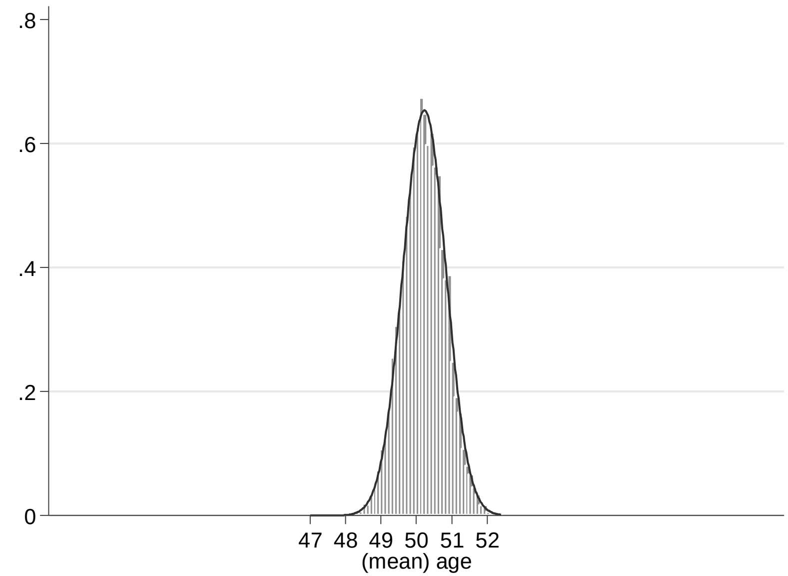





```
. summm age

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
         age |      10000    50.26716    1.906907      43.37      57.24
```

```
. summm age

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
         age |      10000    50.23016    .6103055     47.965     52.399
```



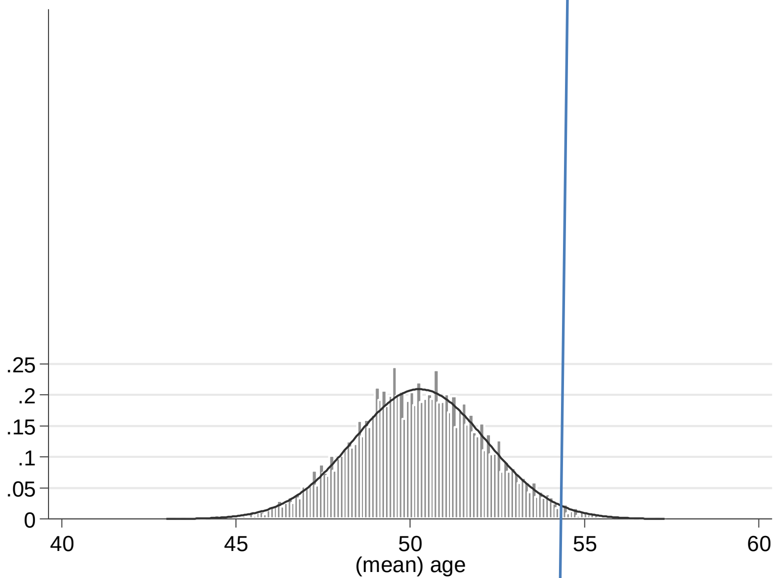


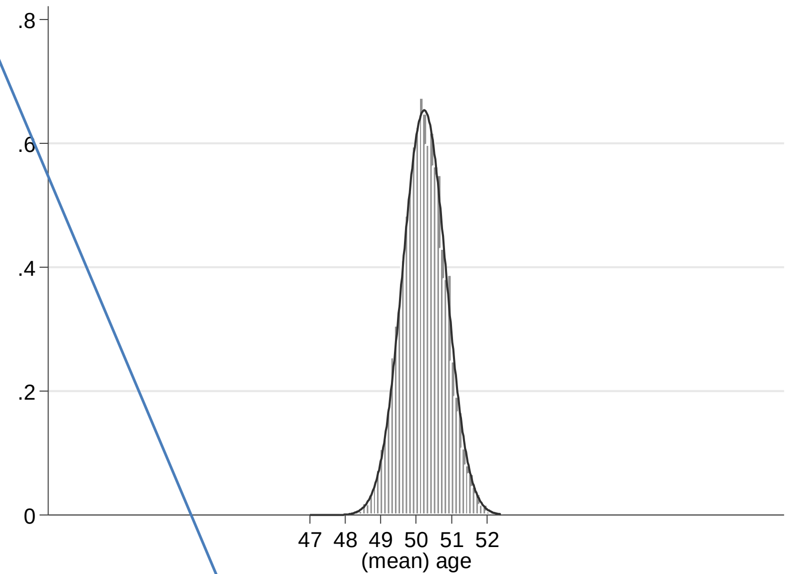

```
. summm age

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
         age |     10000    50.26716    1.906907      43.37      57.24
```

```
. summm age

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
         age |     10000    50.23016    .6103055     47.965     52.399
```

Standard error

Note that 1.906907/.6103055 ≈ √(1,000/100)






```
. summ age

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
         age |     10000    50.26716    1.906907      43.37      57.24
```

```
. summ age

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
         age |     10000    50.23016    .6103055     47.965     52.399
```

Reasoning Backward

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

but want to say something about μ

Reasoning Backward

When you know n, $\overline{\mathrm{X}}$, 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 51.37, with a standard deviation of 18.95, 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 51.37, with a standard deviation of 18.95, 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 51.37, with a standard deviation of 18.95, what can we say about the likely *population* mean?

The most likely value of the population of the mean is 51.37 If we draw repeated samples of 1,000, we expect the standard deviation of those draws (i.e., the standard error) to be:

$$18.95/\sqrt{1{,}000} = 0.60$$

Most important standard errors

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

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

Most important standard errors

“Calc # 1
in computing
“margin of error”

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

Most important standard errors

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

“Diff. of means *t*-tests

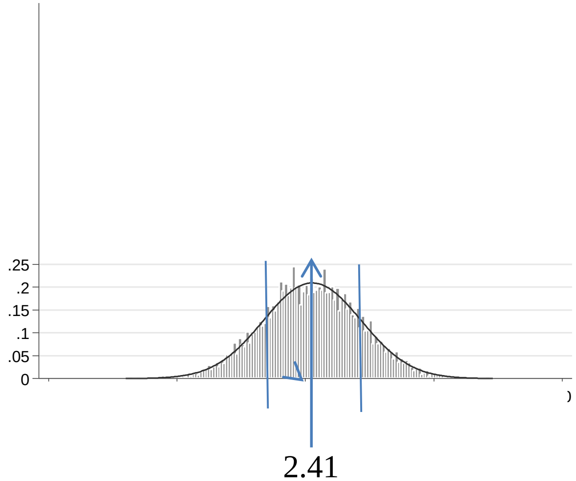
Standard error of a mean

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

Example:
We have drawn a sample of 49,921 American adults, in which they report that, on average, they rate Obama a 2.41 on a four-point approval scale (4 = approve strongly). The standard deviation is 1.22. What is the standard error?

$$1.22 / \sqrt{49{,}921} = 0.01$$

Standard error of a mean



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

Example:

We have drawn a sample of 49,921 American adults, in which they report that, on average, they rate Obama a 2.41 on a four-point approval scale (4 = approve strongly). The standard deviation is 1.22. What is the standard error?

$$1.22/\sqrt{49,921} = 0.01$$

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 49,921 voters nationwide and have run a regression. The dependent variable is the four-point Obama approval scale. The independent variable is a 5-point ideology scale (5 = very liberal.) All the relevant information is on the next slide. What is the standard error of the regression coefficient?

```
. reg obama_approve liberal [aw=V103]
(sum of wgt is   4.9997e+04)

      Source |       SS       df       MS              Number of obs =   49921
-------------+------------------------------           F(  1, 49919) =25710.50
       Model |  25387.7163     1  25387.7163           Prob > F      =  0.0000
    Residual |  49292.2951 49919  .987445563           R-squared     =  0.3400
-------------+------------------------------           Adj R-squared =  0.3399
       Total |  74680.0114 49920  1.49599382           Root MSE      =   .9937

------------------------------------------------------------------------------
obama_appr~e |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     liberal |   .7068479
       _cons |   .3809636
------------------------------------------------------------------------------
```

```
. tabstat liberal [aw=V103],statistics(sd)

    variable |        sd
-------------+----------
     liberal |  1.006351
------------------------
```

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

$$\frac{0.9937}{\sqrt{49920}} \times \frac{1}{1.006} = 0.004421$$

```
. reg obama_approve liberal [aw=V103]
(sum of wgt is   4.9997e+04)

      Source |       SS       df       MS              Number of obs =   49921
-------------+------------------------------           F(  1, 49919) =25710.50
       Model |  25387.7163     1  25387.7163           Prob > F      =  0.0000
    Residual |  49292.2951 49919  .987445563           R-squared     =  0.3400
-------------+------------------------------           Adj R-squared =  0.3399
       Total |  74680.0114 49920  1.49599382           Root MSE      =   .9937

------------------------------------------------------------------------------
obama_appr~e |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     liberal |   .7068479   .0044083
       _cons |   .3809636   .0133936
------------------------------------------------------------------------------
```

```
. tabstat liberal [aw=V103],statistics(sd)

    variable |        sd
-------------+----------
     liberal |  1.006351
------------------------
```

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

$$\frac{0.9937}{\sqrt{49920}} \times \frac{1}{1.006} = 0.004421$$

Factors that reduce the standard error

Overall goodness of fit

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

Sample size

Variance of ind. var.

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 therein

- ci = sample parameter $\pm$

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 therein

- ci = sample parameter ± multiple * sample standard error

Based on the normal curve

Constructing Confidence Intervals

Example:
We have drawn a sample of 49,921 voters nationwide and have run a regression. The dependent variable is the four-point Obama approval scale. The independent variable is a 5-point ideology scale (5 = very liberal.) All the relevant information is on the next slide. What is the 95% confidence interval

ci = sample parameter $\pm$ multiple * sample standard error

Constructing Confidence Intervals

Example:
We have drawn a sample of 49,921 voters nationwide and have run a regression. The dependent variable is the four-point Obama approval scale. The independent variable is a 5-point ideology scale (5 = very liberal.) All the relevant information is on the next slide. What is the 95% confidence interval

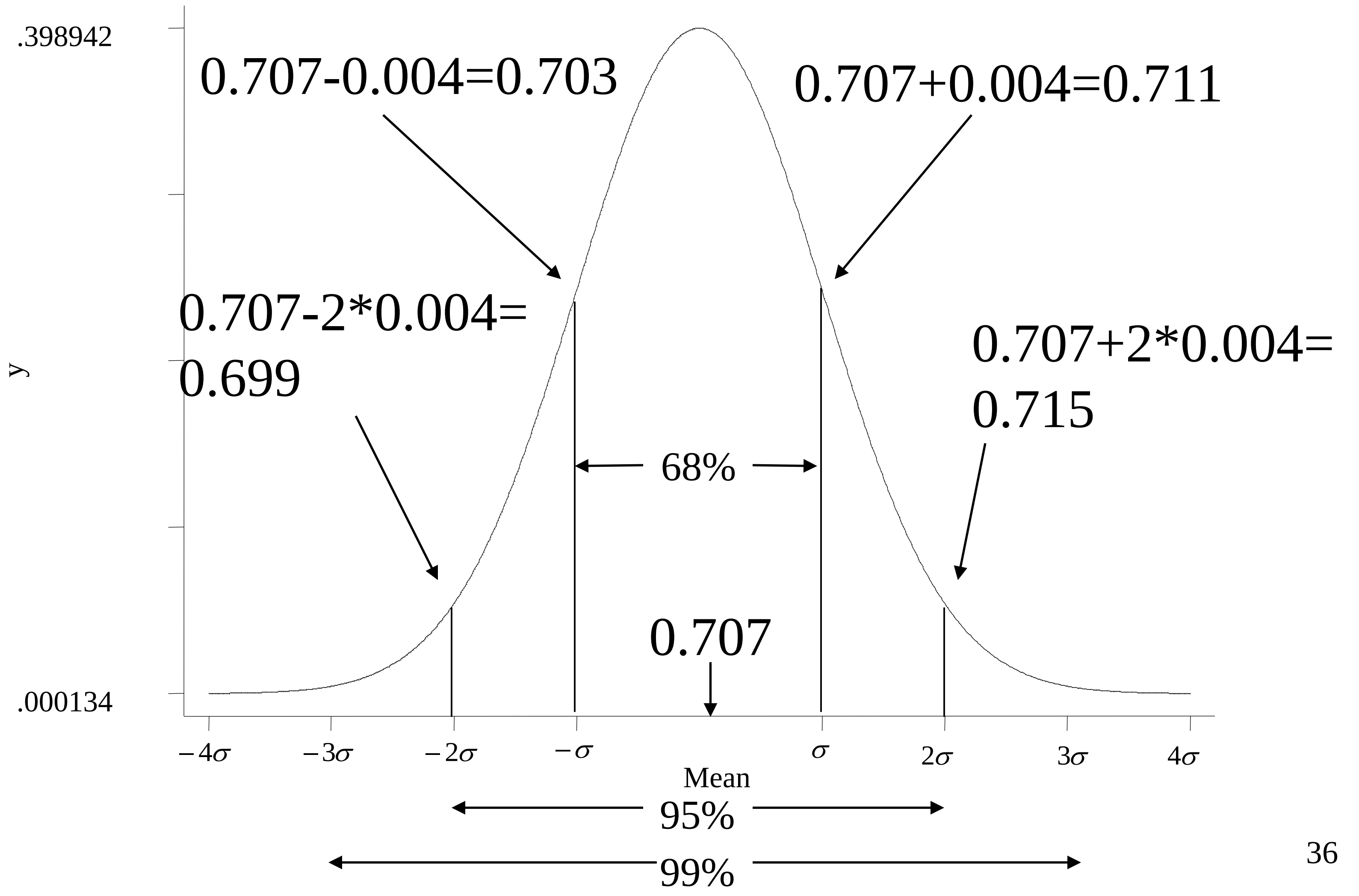
ci = sample parameter $\underline{+}$ multiple * sample standard error

0.707

$$\frac{s.e.r.}{\sqrt{n-1}} \times \frac{1}{s_x} = \frac{0.9937}{\sqrt{49920}} \times \frac{1}{1.006} = 0.004421$$

N = 49,921; coeff. = 0.707; s.e. = 0.004421 (round to 0.004)

The Picture



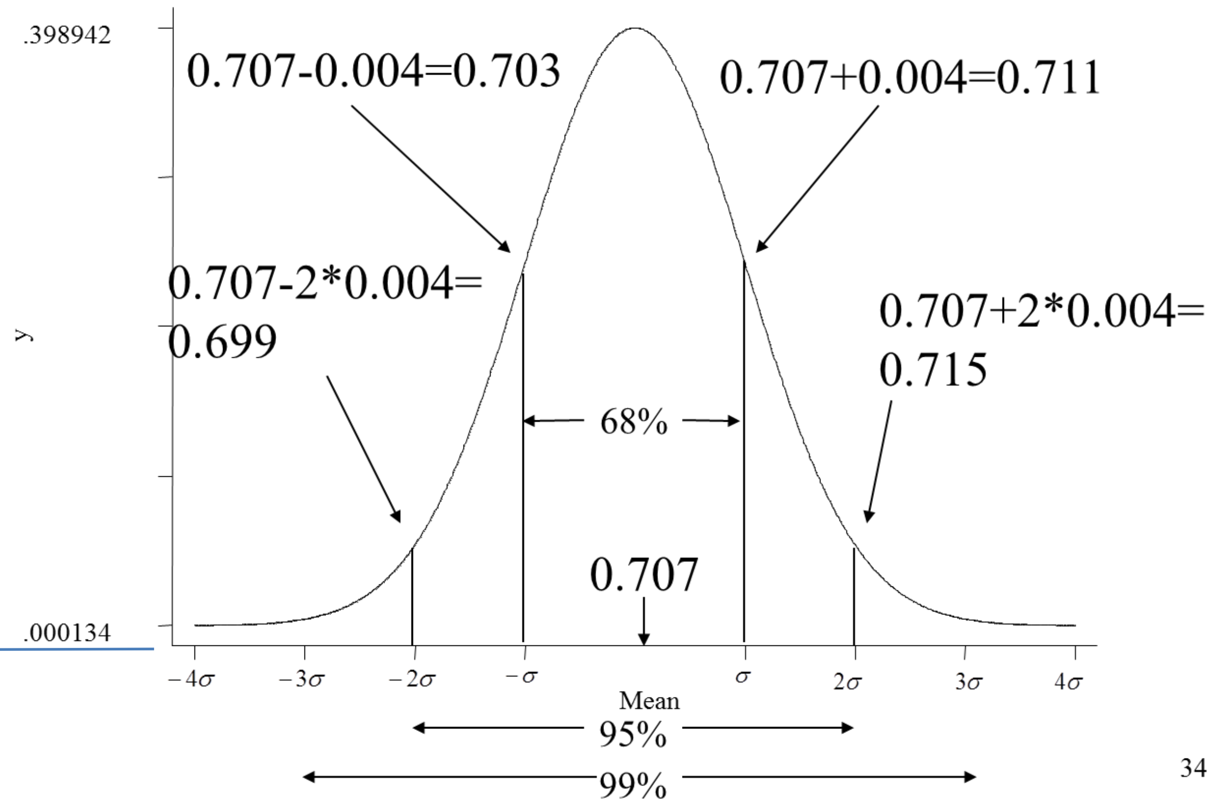
Confidence Intervals for coefficient

- 68% confidence interval = 0.707 ± 0.004 = [0.703 to 0.711]
- 95% confidence interval = $0.707 \pm 2*0.004$ = [0.699 to 0.715]
- 99% confidence interval = $0.707 \pm 3*.004$ = [0.695 to 0.719]

```
. reg obama_approve liberal [aw=V103]
(sum of wgt is   4.9997e+04)

      Source |       SS       df       MS              Number of obs =   49921
-------------+------------------------------           F(  1, 49919) =25710.50
       Model |  25387.7163     1  25387.7163           Prob > F      =  0.0000
    Residual |  49292.2951 49919  .987445563           R-squared     =  0.3400
-------------+------------------------------           Adj R-squared =  0.3399
       Total |  74680.0114 49920  1.49599382           Root MSE      =   .9937

------------------------------------------------------------------------------
obama_appr~e |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     liberal |   .7068479   .0044083  [redacted]          .6982076    .7154882
       _cons |   .3809636   .0133936                      .3547119    .4072153
------------------------------------------------------------------------------
```

- 68% confidence interval = 0.707±0.004 = [0.703 to 0.711]
- 95% confidence interval = 0.707±2*0.004 = [0.699 to 0.715]
- 99% confidence interval = 0.707±3*.004 = [0.695 to 0.719]

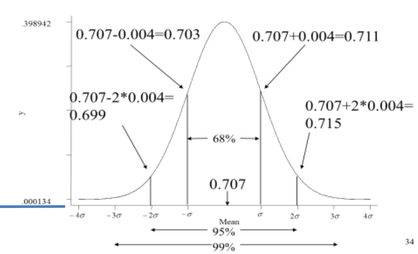
The final question:
How far is the coefficient from some target amount (usually zero)?



Zero is
down here
somewhere

```
. reg obama_approve liberal [aw=V103]
(sum of wgt is   4.9997e+04)

      Source |       SS       df       MS              Number of obs =   49921
-------------+------------------------------           F(  1, 49919) =25710.50
       Model |  25387.7163     1  25387.7163           Prob > F      =  0.0000
    Residual |  49292.2951 49919  .987445563           R-squared     =  0.3400
-------------+------------------------------           Adj R-squared =  0.3399
       Total |  74680.0114 49920  1.49599382           Root MSE      =   .9937

------------------------------------------------------------------------------
obama_appr~e |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     liberal |   .7068479   .0044083   160.34   0.000     .6982076    .7154882
       _cons |   .3809636   .0133936    28.44   0.000     .3547119    .4072153
------------------------------------------------------------------------------
```

$$t = \frac{coeff.-target\ value}{s.e.} = \frac{0.707-0}{.00441} = 160.3$$



Zero is down here somewhere at 160.3σ

Residual Vote Multivariate Analysis, Presidential, Gubernatorial, and Senatorial Elections, 1988–2000

	President	President
Equipment effects:		
Punch card	.0082	.0077
	(.0015)	(.0005)
Lever machine	Excluded equip. category	Excluded equip. category
Paper	−.014	−.0012
	(.002)	(.0014)
Optical scan	−.0045	.00071
	(.0014)	(.00070)
Electronic (DRE)	.0022	.0080
	(.0015)	(.0010)
Shift in tech.	.0010	.00005
	(.0007)	(.00067)
Log (turnout)	.0095	−.0004
	(.0026)	(.0001)
Gov. or Sen. on ballot	−.0011	−.003
	(.0007)	(.001)
Senator	—	—
Percent Over 65	—	.047
		(.008)
Percent 18–24	—	−.012
		(.009)
Percent White	—	−.030
		(.002)
Percent Hispanic	—	.011
		(.004)
Median Income (10,000s)	—	−.002
		(.001)
Constant	−.11	.025
	(.03)	(.002)
N	8,982	8,982
R^2	.79	.14
Fixed effect: (not shown)	Year × State County	Year × State
Number of categories	3,346	—

More normal t-statistics

$$t = \frac{-0.0045}{0.0014} = -3.21$$

Associated with a 99.9% c.i.

Residual Vote Multivariate Analysis, Presidential, Gubernatorial, and Senatorial Elections, 1988–2000

More normal
t-statistics

	President	President
Equipment effects:		
Punch card	.0082	.0077
	(.0015)	(.0005)
Lever machine	Excluded equip. category	Excluded equip. category
Paper	–.014	–.0012
	(.002)	(.0014)
Optical scan	–.0045	.00071
	(.0014)	(.00070)
Electronic (DRE)	.0022	.0080
	(.0015)	(.0010)
Shift in tech.	.0010	.00005
	(.0007)	(.00067)
Log (turnout)	.0095	–.0004
	(.0026)	(.0001)
Gov. or Sen. on ballot	–.0011	–.003
	(.0007)	(.001)
Senator	—	—
Percent Over 65	—	.047
		(.008)
Percent 18–24	—	–.012
		(.009)
Percent White	—	–.030
		(.002)
Percent Hispanic	—	.011
		(.004)
Median Income (10,000s)	—	–.002
		(.001)
Constant	–.11	.025
	(.03)	(.002)
N	8,982	8,982
R^2	.79	.14
Fixed effect: (not shown)	Year × State County	Year × State
Number of categories	3,346	—

$$t = \frac{-0.0022}{0.0015} = 1.47$$

Associated with a 85.8% c.i.

A word about standard errors with multiple regression

$$s.e._{\beta_k} = \frac{s.e.r.}{\sqrt{(1 - R^2{}_{X_k G_k})(n-1)}} \times \frac{1}{s_{x_k}}$$

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

A word about standard errors with multiple regression

If the other independent variables are orthogonal to X_k

$$s.e._{\beta_k} = \frac{s.e.r.}{\sqrt{(1 - R^2{}_{X_k G_k})(n-1)}} \times \frac{1}{s_{x_k}}$$

0

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

A word about standard errors with multiple regression

If the other independent variables are orthogonal to X_k

$$s.e._{\beta_k} = \frac{s.e.r.}{\sqrt{(1 - R^2{}_{X_k G_k})(n-1)}} \times \frac{1}{s_{x_k}}$$

0

But remember…

$$s.e.r. = \frac{\sum(y_i - \widehat{y_i})^2}{n - m - 1}$$

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

A word about standard errors with multiple regression

If the other independent variables are perfectly collinear with X_k

$$s.e._{\beta_k} = \frac{s.e.r.}{\sqrt{(1 - \underbrace{R^2_{X_k G_k}}_{1})(n-1)}} \times \frac{1}{s_{x_k}}$$

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