Lecture 23: Chapter 13, Section 1
Two Quantitative Variables
Inference for Regression

- Regression for Sample vs. Population
- Population Model; Parameters and Estimates
- Regression Hypotheses
- Test about Slope; Interpreting Output
- Confidence Interval for Slope

Looking Back: *Review*

- **4 Stages of Statistics**
  - Data Production (discussed in Lectures 1-3)
  - Displaying and Summarizing (Lectures 3-8)
  - Probability (discussed in Lectures 9-14)
  - Statistical Inference
    - 1 categorical (discussed in Lectures 14-16)
    - 1 quantitative (discussed in Lectures 16-18)
    - cat and quan: paired, 2-sample, several-sample (Lectures 19-21)
    - 2 categorical (discussed in Lectures 21-22)
    - 2 quantitative

Correlation and Regression *(Review)*

- Relationship between 2 quantitative variables
  - Display with scatterplot
  - Summarize:
    - Form: linear or curved
    - Direction: positive or negative
    - Strength: strong, moderate, weak

    If form is linear, correlation r tells direction and strength.

    Also, equation of least squares regression line lets us predict a response $\hat{y}$ for any explanatory value x.

Regression Line and Residuals *(Review)*

Summarize linear relationship between explanatory (*x*) and response (*y*) values with line $\hat{y} = b_0 + b_1 x$ minimizing sum of squared prediction errors $y_i - \hat{y}_i$ (called *residuals*). Typical residual size is

$$s = \sqrt{\frac{(y_1 - \hat{y}_1)^2 + \cdots + (y_n - \hat{y}_n)^2}{n-2}}$$

- **Slope:** predicted change in response *y* for every unit increase in explanatory value *x*
- **Intercept:** predicted response for *x*=0

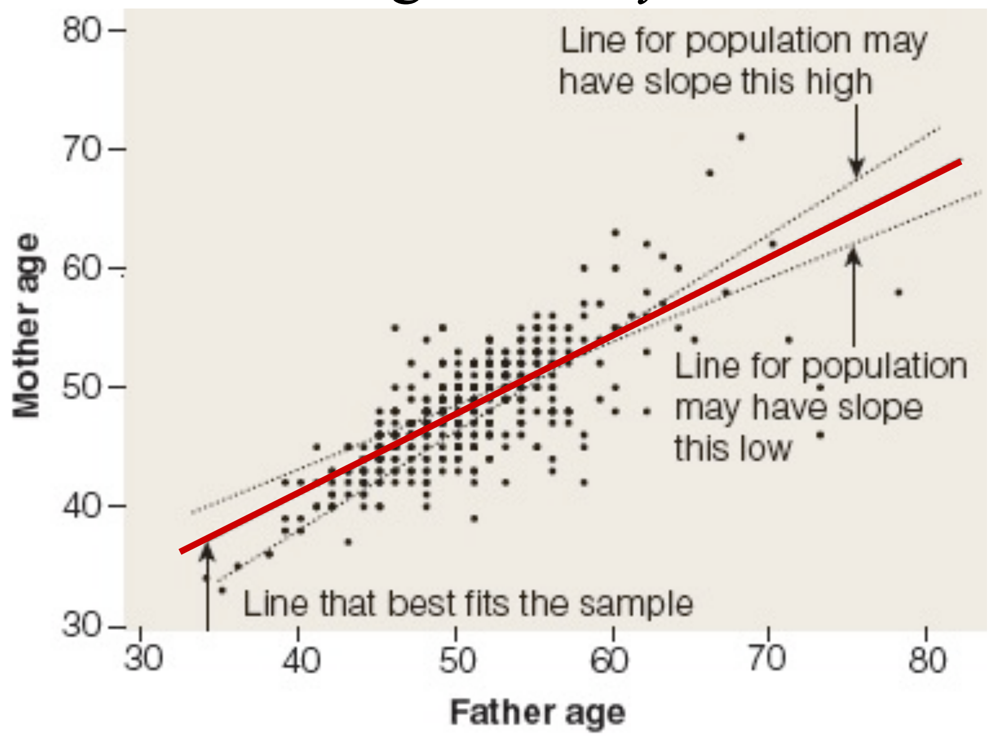
Note: this is the line that best fits the ***sampled*** points.

Regression for Sample vs. Population

- Can find line that best fits the *sample.*
- What does it tell about line that best fits *population*?

Example: *Slope for Sample, Population*

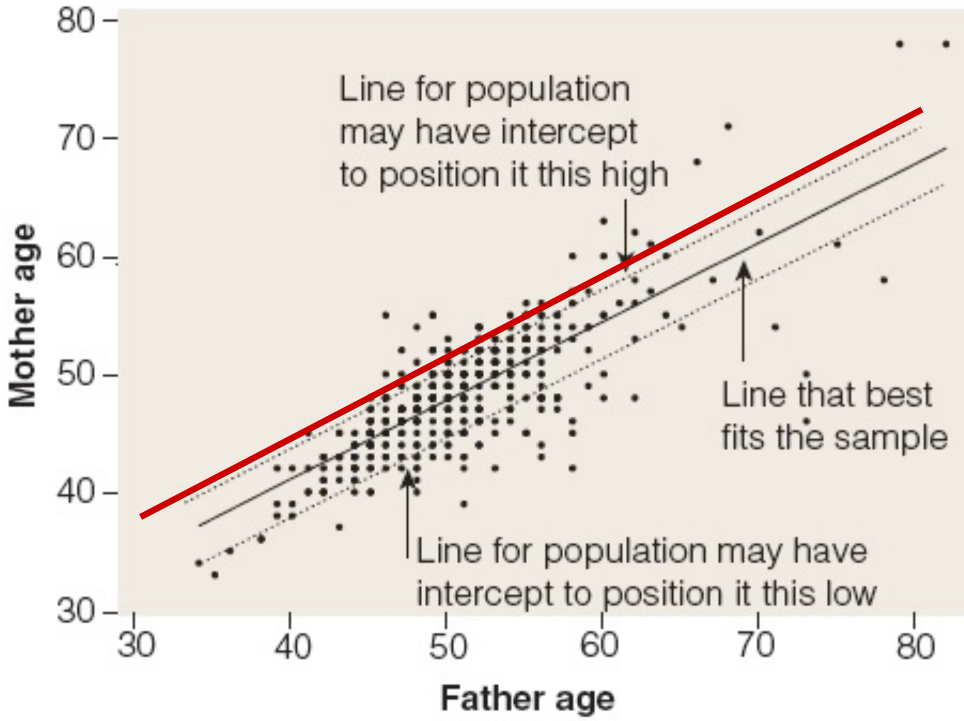
- **Background**: Parent ages have $\hat{y} = 14.54+0.666x$, $s = 3.3$.



- **Question:** Is 0.666 the slope of the line that best fits relationship for *all* students' parents ages?
- **Response:** Slope β_1 of best line for *all* parents is

Example: *Intercept for Sample, Population*

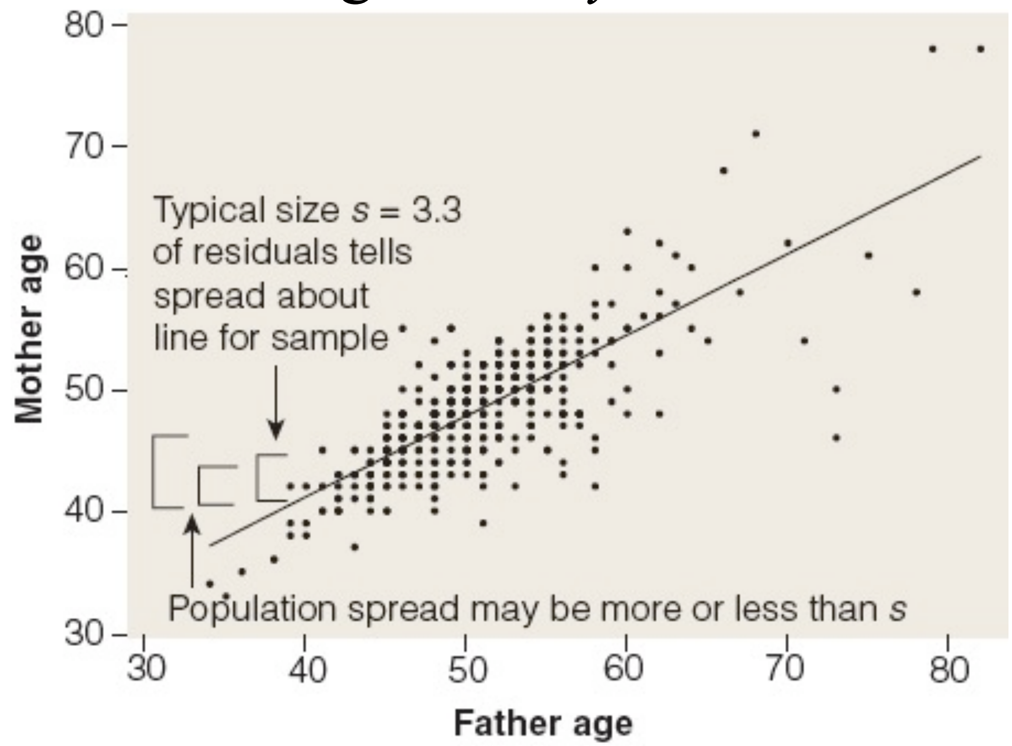
- **Background**: Parent ages have $\hat{y} = 14.54 + 0.666x$, $s = 3.3$.



- **Question:** Is 14.54 the intercept of the line that best fits relationship for *all* students' parents ages?
- **Response:** Intercept β_o of best line for *all* parents is

Example: *Prediction Error for Sample, Pop.*

- **Background**: Parent ages have $\hat{y} = 14.54+0.666x$, $s = 3.3$.



- **Question:** Is 3.3 the typical prediction error size for the line that relates ages of *all* students' parents?
- **Response:** Typical residual size for best line for *all* parents is ______________________________

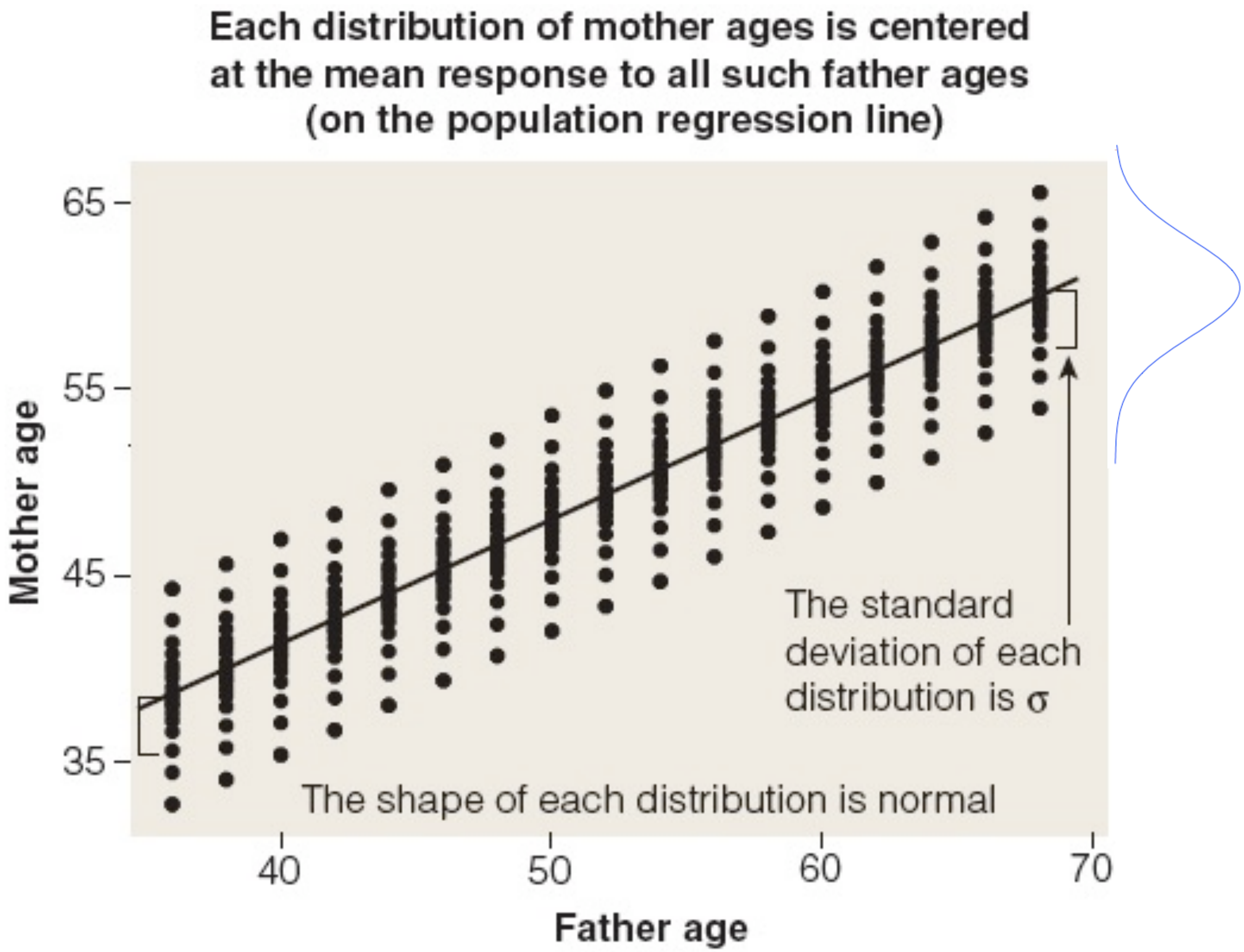
Notation; Population Model; Estimates

σ : **typical residual size** for line best fitting linear relationship in population.

$\mu_y = \beta_o + \beta_1 x$: **population mean response** to any *x*. Responses vary normally about μ_y with standard deviation σ

Parameter	Estimate
β_o	b_o
β_1	b_1
σ	s

Population Model



Estimates

Parameter	Estimate
β_o	b_o
β_1	b_1
σ	s

- Intercept and spread: point estimates suffice.
- Slope is focus of regression inference (hypothesis test, sometimes confidence interval).

Regression Hypotheses

- $H_o : \beta_1 = 0 \rightarrow \mu_y = \beta_o + \beta_1 x$

  → no population relationship between *x* and *y*

- $H_a : \beta_1 \begin{Bmatrix} > \\ < \\ \neq \end{Bmatrix} 0$

  → *x* and *y* *are* related for population (and relationship is positive if >, negative if <)

Example: *Point Estimates and Test about Slope*

- **Background**: Consider parent age regression:

```
The regression equation is
MotherAge = 14.5 + 0.666 FatherAge
431 cases used 15 cases contain missing values
Predictor        Coef    SE Coef        T       P
Constant       14.542      1.317    11.05   0.000
FatherAge     0.66576    0.02571    25.89   0.000
S = 3.288      R-Sq = 61.0%     R-Sq(adj) = 60.9%
```

- **Questions:** What are parameters of interest and accompanying estimates? What hypotheses will we test?
- **Responses:** For $\mu_y = \beta_o + \beta_1 x$, estimate
  - Parameter____with ________
  - Parameter____with ________
  - Parameter____with ________
  - Test H_o : vs. H_a :

Suspect ________ relationship.

Practice: 13.9 p.648

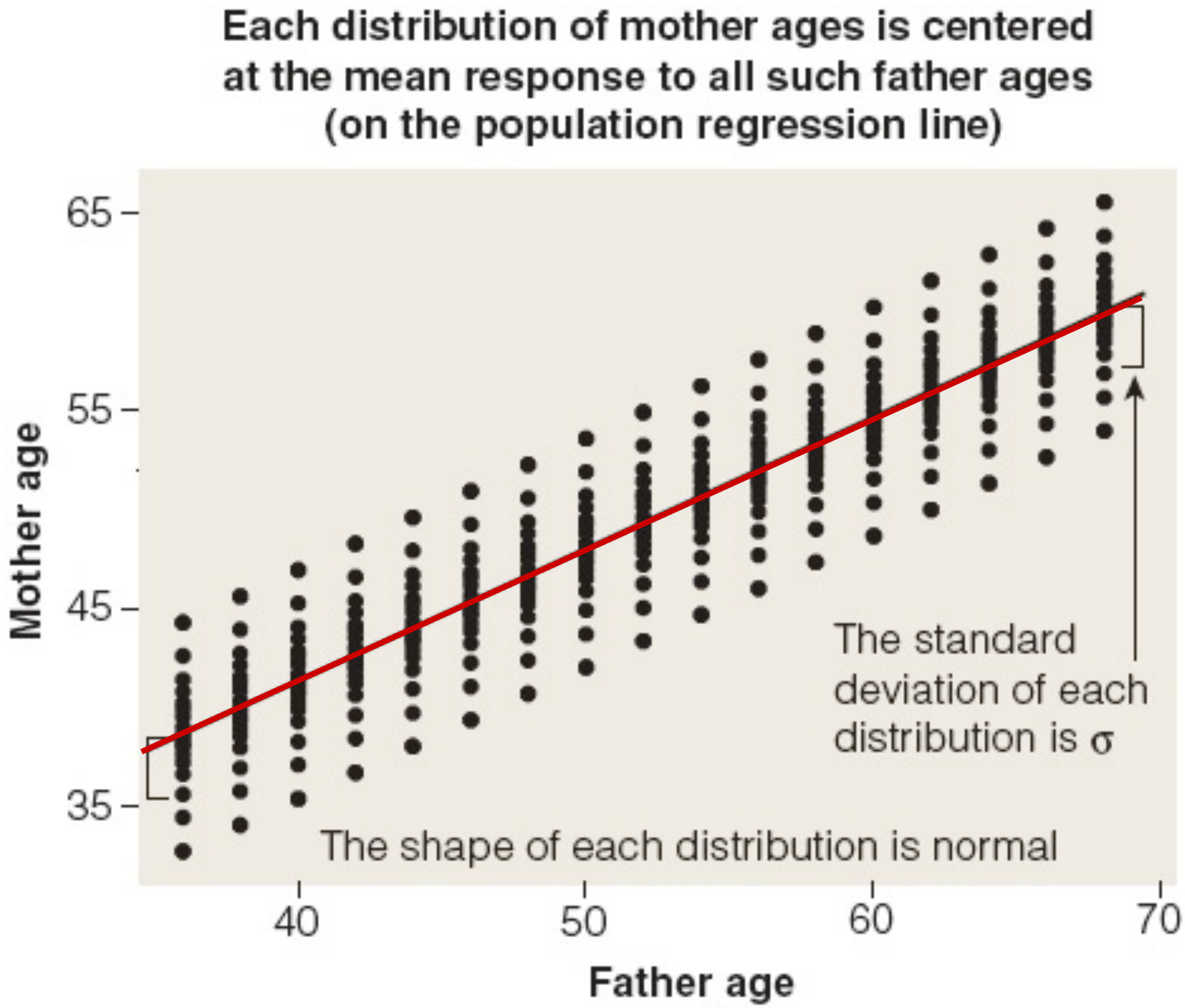
Key to Solving Inference Problems *(Review)*

(1 quantitative variable) For a given population mean μ, standard deviation σ, and sample size *n*, needed to find probability of sample mean $\bar{X}$ in a certain range:

Needed to know **sampling distribution** of $\bar{X}$ in order to perform inference about μ.

Now, to perform inference about β_1, need to know sampling distribution of b_1.

Slopes b_1 from Random Samples Vary

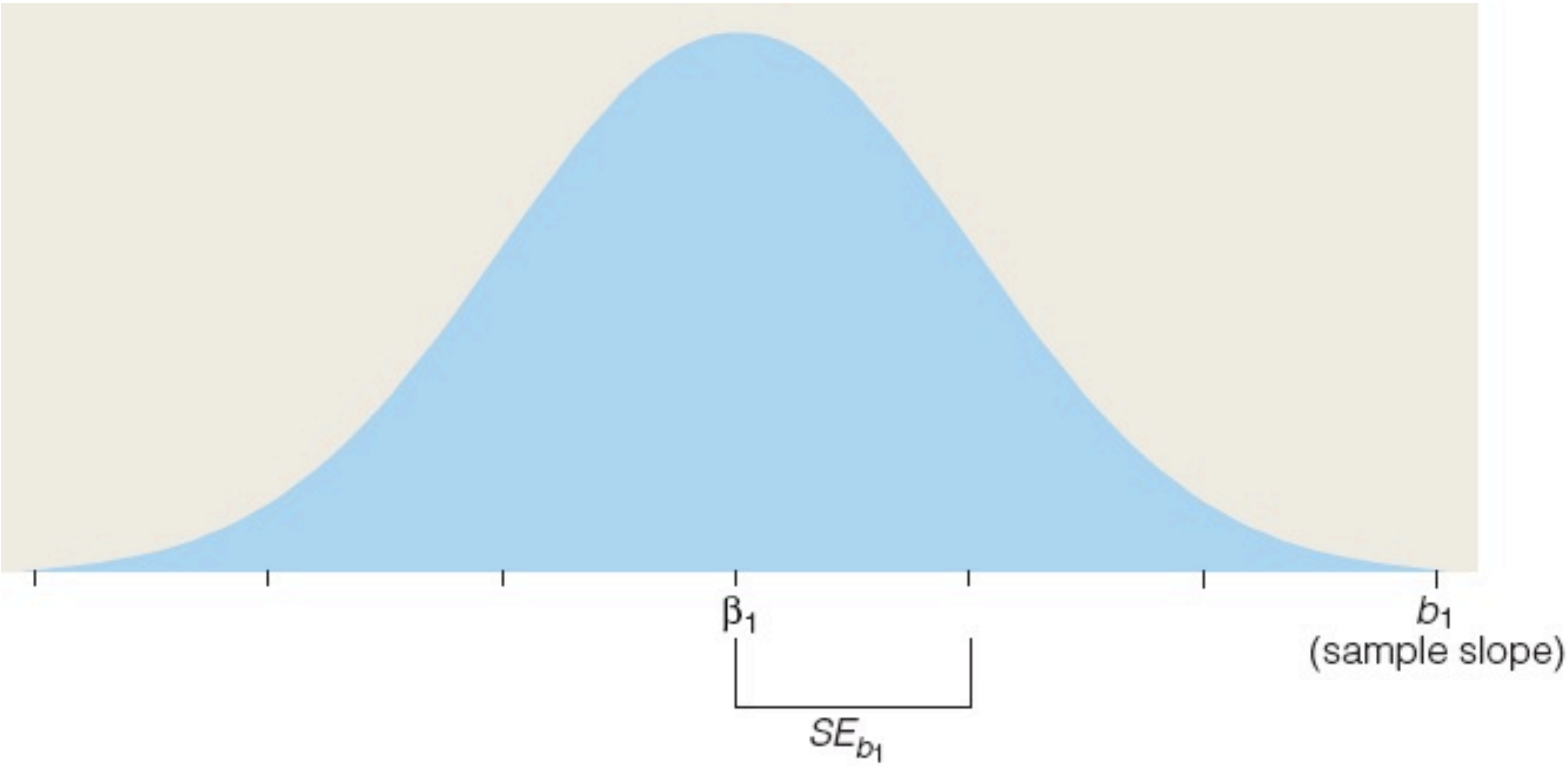


Distribution of Sample Slope

As a random variable, sample slope b_1 has

- Mean β_1
- s.d. $\approx SE_{b_1} = \frac{\boxed{s}}{\sqrt{(x_1 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}}$
  - Residuals large→slope hard to pinpoint
  - Residuals small→slope easy to pinpoint
- Shape approximately normal if responses vary normally about line, or *n* large

Distribution of Sample Slope



Distribution of Standardized Sample Slope

Standardize b_1 to $t = \frac{b_1 - \beta_1}{SE_{b_1}}$

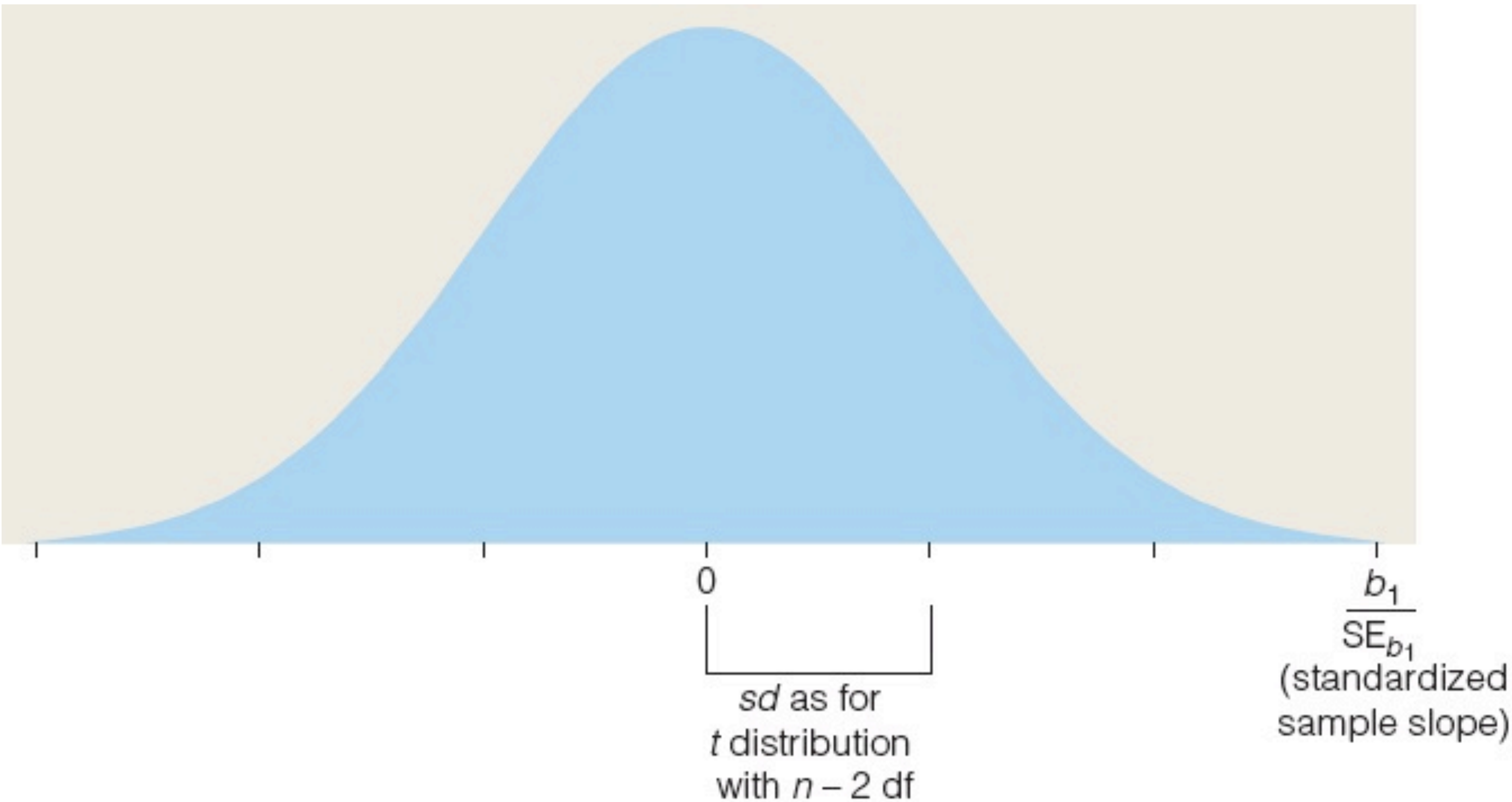
$= \frac{b_1 - 0}{SE_{b_1}}$ if H_0 is true.

For large enough *n*, ***t*** follows *t* distribution with *n*-2 degrees of freedom.

- b_1 close to 0→*t* not large→*P*-value not small
- b_1 far from 0→*t* large→*P*-value small

Sample slope far from 0 gives evidence to reject Ho, conclude population slope not 0.

Distribution of Standardized Sample Slope



Example: *Regression Output (Review)*

- **Background**: Regression of mom and dad ages:

```
The regression equation is
MotherAge = 14.5 + 0.666 FatherAge
431 cases used 15 cases contain missing values
Predictor          Coef     SE Coef          T        P
Constant         14.542       1.317      11.05    0.000
FatherAge       0.66576     0.02571      25.89    0.000
S = 3.288          R-Sq = 61.0%      R-Sq(adj) = 60.9%
```

- **Question:** What does the output tell about the relationship between mother' and fathers' ages in the sample?
- **Response:**
  - Line_____________________best fits sample (slope pos).
  - Sample relationship_______: r =__________________
  - Typical size of prediction errors for sample is ________

Practice: 13.2c,d,l p.646

Example: *Regression Inference Output*

- **Background**: Regression of 431 parent ages:

Predictor	Coef	SE Coef	T	P
Constant	14.542	1.317	11.05	0.000
FatherAge	0.66576	0.02571	25.89	0.000

S = 3.288   R-Sq = 61.0%   R-Sq(adj) = 60.9%

- **Question:** What does the output tell about the relationship between mother' and fathers' ages in the population?
- **Response:** To test $H_o : \beta_1 = 0$ vs. $H_a : \beta_1 > 0$ focus on _____line of numbers (about slope, not intercept)
  - Estimate for slope of line best fitting population:______
  - Standard error of sample slope:____________________
  - Stan. sample slope:______________________________
  - *P*-value: _________ = 0.000 where *t* has df =________
  - Reject H_0? _____ Variables related in population?_____

Practice: 13.8d-e p.648

Strength of Relationship or of Evidence

- Can have weak/strong evidence of weak/strong relationship.
- Correlation r tells strength of relationship (observed in sample)
  - $|r|$ close to 1 → relationship is strong
- P-value tells strength of evidence that variables are related in population.
  - P-value close to 0 → evidence is strong

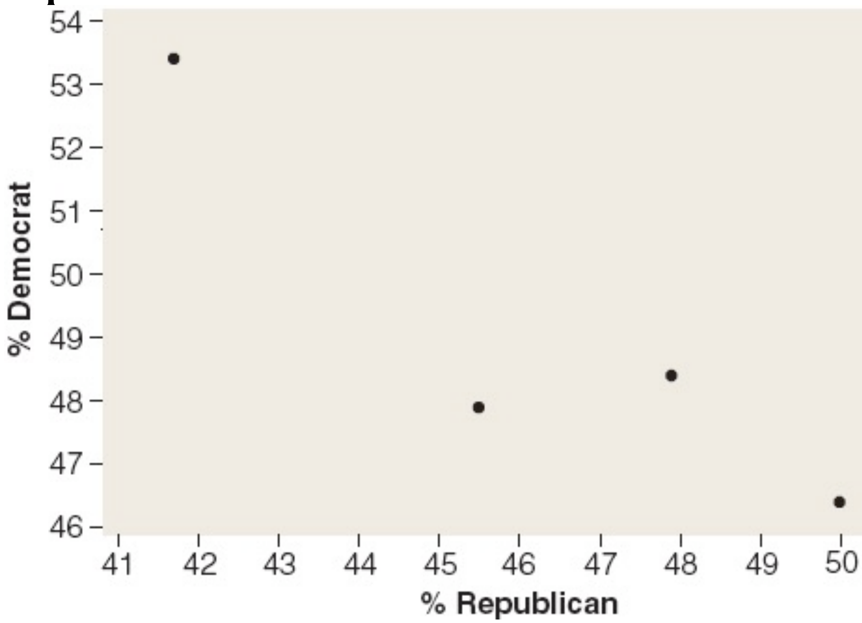
Example: *Strength of Relationship, Evidence*

- **Background**: Regression of students' mothers' on fathers' ages had r=+0.78, p=0.000.
- **Question:** What do these tell us?
- **Response:**
  - r fairly close to 1→ ______________________________
  - P-value 0.000→

    ____
  - We have ____________ evidence of a ________________ relationship between students' mothers' and fathers' ages in general.

Example: *Strength of Evidence; Small Sample*

- **Background**: % voting Dem vs. % voting Rep for 4 states in 2000 presidential election has $r = -0.922$, P-value 0.078.

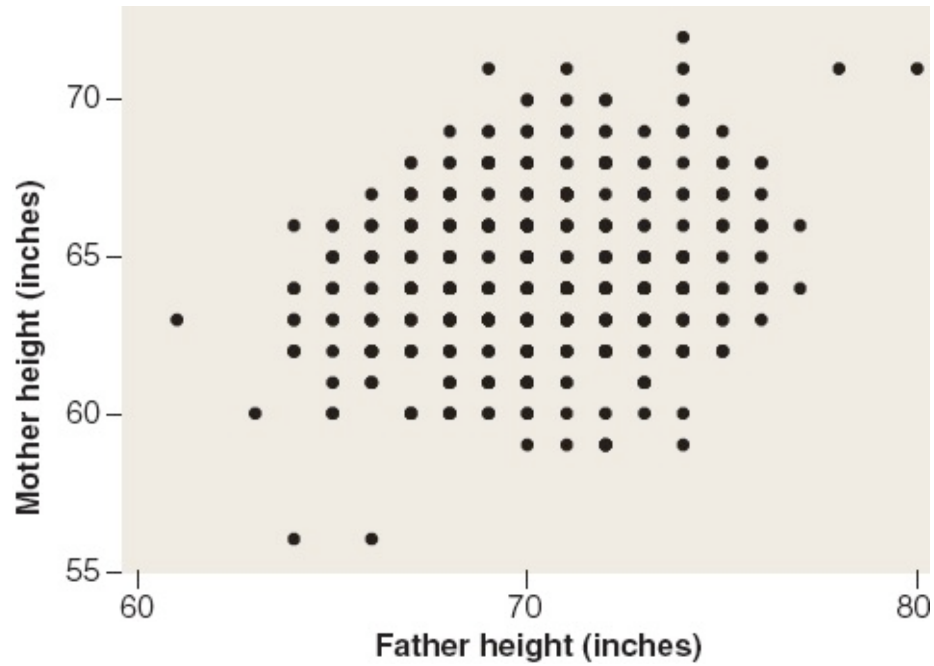


- **Question:** What do these tell us?
- **Response:** We have _______ evidence (due to _____________) of a_________ relationship in the population of states.

Practice: 13.10 p.649

Example: *Strength of Evidence; Large Sample*

- **Background**: Hts of moms vs. hts of dads have *r* =+0.225, *P*-value 0.000.



- **Question:** What do these tell us?
- **Response:** There is ________ evidence (due to ____________) of a ________ relationship in the population.

Distribution of Sample Slope *(Review)*

As a random variable, sample slope b_1 has

- Mean β_1
- s.d. $\approx SE_{b_1} = \frac{s}{\sqrt{(x_1-\bar{x})^2+\cdots+(x_n-\bar{x})^2}}$
- Shape approximately normal if responses vary normally about line, or *n* large

To construct confidence interval for unknown population slope β_1 use b_1 as estimate, SE_{b_1} as estimated s.d., and *t* multiplier with *n*-2 df.

Confidence Interval for Slope

Confidence interval for β_1 is

$$b_1 \pm multiplier(SE_{b_1})$$

where multiplier is from *t* dist. with *n*-2 df.

If *n* is large, 95% confidence interval is

$$b_1 \pm 2(SE_{b_1}).$$

Example: *Confidence Interval for Slope*

- **Background**: Regression of 431 parent ages:

Predictor	Coef	SE Coef	T	P
Constant	14.542	1.317	11.05	0.000
FatherAge	0.66576	0.02571	25.89	0.000

S = 3.288   R-Sq = 61.0%   R-Sq(adj) = 60.9%

- **Question:** What is an approximate 95% confidence interval for the slope of the line relating mother's age and father's age for all students?
- **Response:** Use multiplier ______________________

We're 95% confident that for population of age pairs, if a father is 1 year older than another father, the mother is on average between_____ and_____ years older.

Note: Interval ______________________ ←→Rejected Ho.

Lecture Summary
(Inference for Quan→Quan: Regression)

- Regression for sample vs. population
  - Slope, intercept, sample size
- Regression hypotheses
- Test about slope
  - Distribution of sample slope
  - Distribution of standardized sample slope
- Regression inference output
  - Strength of relationship, strength of evidence
- Confidence interval for slope