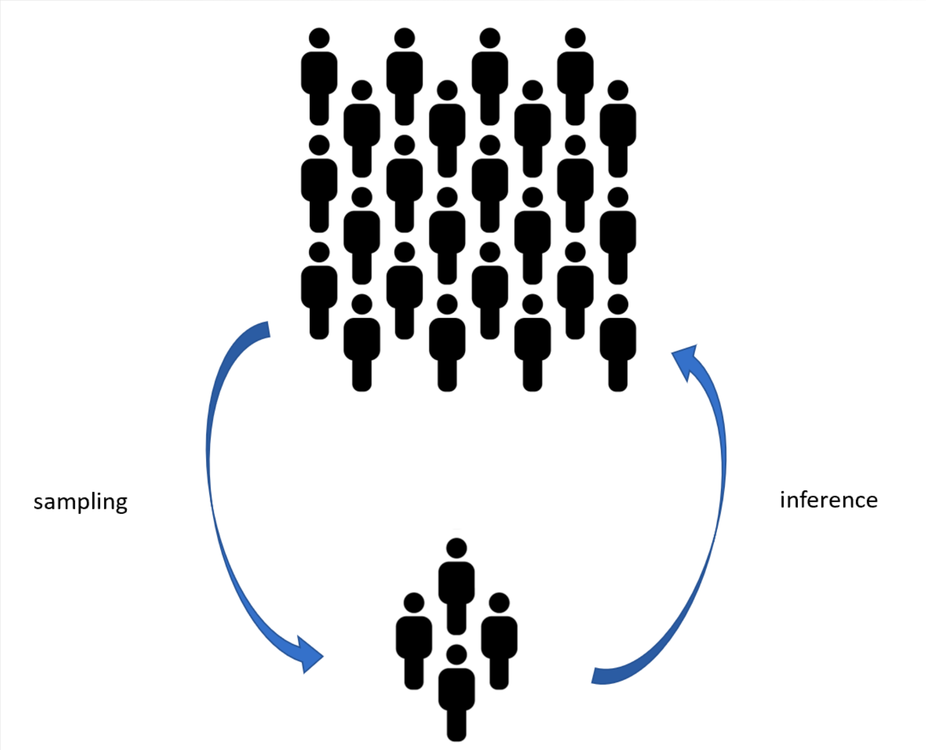
3 – REVIEW OF STATISTICS

University *of* Massachusetts Amherst BE REVOLUTIONARY™

THE PLAN

1. Estimating the Population Mean
2. Hypothesis Tests
3. Confidence Intervals
4. Testing Differences between Means
5. Scatterplots and Sample Correlation

WHAT DOES STATISTICS DO?



- Learn about a population by analyzing a random sample.

1. Estimation
2. Hypothesis Testing
3. Confidence Intervals

3.1 ESTIMATING THE POPULATION MEAN

University of Massachusetts Amherst

ESTIMATORS

- **Estimator**: a “best guess” about a population parameter, that can be calculated from sample.

What makes an estimator “good”?

- **Unbiasedness:** $E(\hat{\mu}_Y) = \mu_Y$
- **Consistency:** $\hat{\mu}_Y \xrightarrow{\text{p}} \mu_Y$
- **Efficiency**: $var(\hat{\mu}_Y)$ smaller rather than larger.

$\bar{Y}$ AS AN ESTIMATOR OF μ_Y

- Sample average: $\bar{Y} = \frac{1}{n}(Y_1 + Y_2 + \cdots + Y_n) = \frac{1}{n}\sum_{i=1}^{n} Y_i$

1. $E(\bar{Y}) = \mu_Y$
   1. because $Y_1, Y_{\ldots}, Y_N$ are i.i.d.
2. $\bar{Y} \xrightarrow{p} \mu_Y$ (law of large numbers)
   - because Law of Large Numbers
3. $var(\bar{Y}) < var(\hat{\mu}_Y)$
   - where $\hat{\mu}_Y$=every other linear estimator of μ_Y

$\bar{Y}$ is BLUE

$\bar{Y}$ AS A LEAST SQUARES ESTIMATOR

- *Least square estimator*: the one that minimizes

$$\sum_{i=1}^{n} (Y_i - m)^2$$

- Solution:

$$m = \frac{1}{n}\sum_{i=1}^{n} Y_i = \bar{Y}$$

- $\bar{Y}$ is the *least squares estimator* of μ_Y

$\bar{Y}$ AS A LEAST SQUARES ESTIMATOR: PROOF

$$\min_{m} \sum_{i=1}^{n} (Y_i - m)^2$$

$$\frac{d}{dm} \sum_{i=1}^{n} (Y_i - m)^2 = 2 \sum_{i=1}^{n} (Y_i - m) = 2 \sum_{i=1}^{n} Y_i - 2nm = 0$$

$$\sum_{i=1}^{n} Y_i - nm = 0 \rightarrow m = \frac{1}{n} \sum_{i=1}^{n} Y_i = \bar{Y}$$

IMPORTANCE OF RANDOM SAMPLING

- We are assuming $Y_1, \ldots, Y_n$ are i.i.d., as in random sampling.
- If sampling is not random, $\bar{Y}$ might be biased.
    - $E(\bar{Y}) \neq \mu_Y$
- Is this why pollsters were wrong about Trump in 2016?

EXAMPLE: THE US CENSUS

- US Constitution: count the whole US population every 10 years.
- But some individuals will go undetected
  - Especially minorities, immigrants, poorer families.
- Solution: extrapolate figures for those not counted
  - Democrats like extrapolation, Republicans don't.

3.2 HYPOTESIS TESTS

University *of* Massachusetts Amherst

HYPOTHESIS TESTS: KEY IDEA

- Hypothesis about population parameters:
  - Do average hourly earnings of recent graduates equal 20$/hour?
  - Has more than 70% of the US population been covid-vaccinated?
  - Did the average hourly wage increase in the last year?

- Null hypothesis:

$$H_0 : E(Y) = \mu_{Y,0}$$

- Alternative hypothesis:

$$H_1 : E(Y) \neq \mu_{Y,0}$$

HYPOTHESIS TESTS: P-VALUES

- Your null hypothesis is $H_0 : E(Y) = 20$
- What if in your sample $\bar{Y} = 22.64$?
- **p-value** = $Pr_{H_0}[|\bar{Y} - \mu_{Y,0}| > |\bar{Y}^{act} - \mu_{Y,0}|]$
- Low p-value → null hypothesis is *probably* wrong.
- High p-value → *cannot reject* the null hypothesis.

HOW TO CALCULATE THE P-VALUE

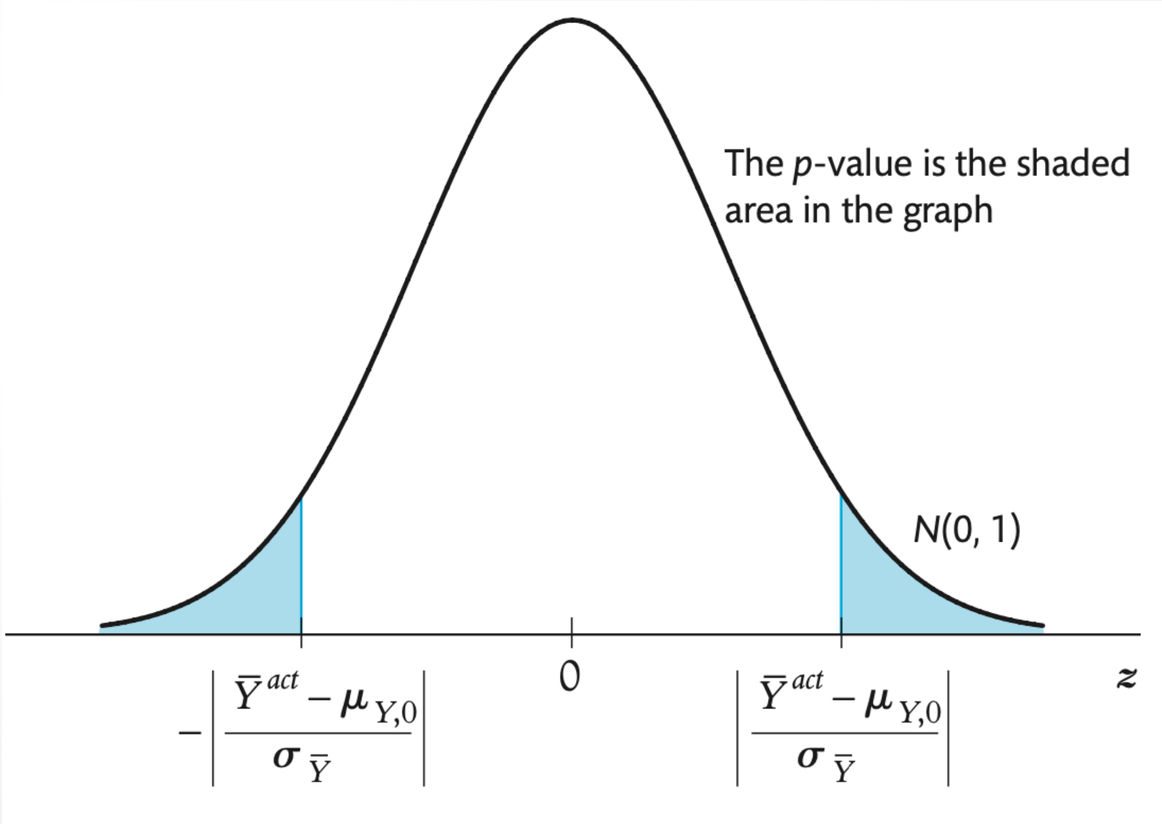
- We need the sampling distribution of $\bar{Y}$ under the null hypothesis

- With large *n*, assuming H_0 true:

$$\bar{Y} \sim N(\mu_{Y,0}, \sigma_{\bar{Y}}^2)$$

- → $\frac{\bar{Y}-\mu_{Y,0}}{\sigma_{\bar{Y}}} \sim N(0,1)$

- P-value = probability that a N(0,1) RV falls as far as $\left|\frac{\bar{Y}^{act}-\mu_{Y,0}}{\sigma_{\bar{Y}}}\right|$ from zero.



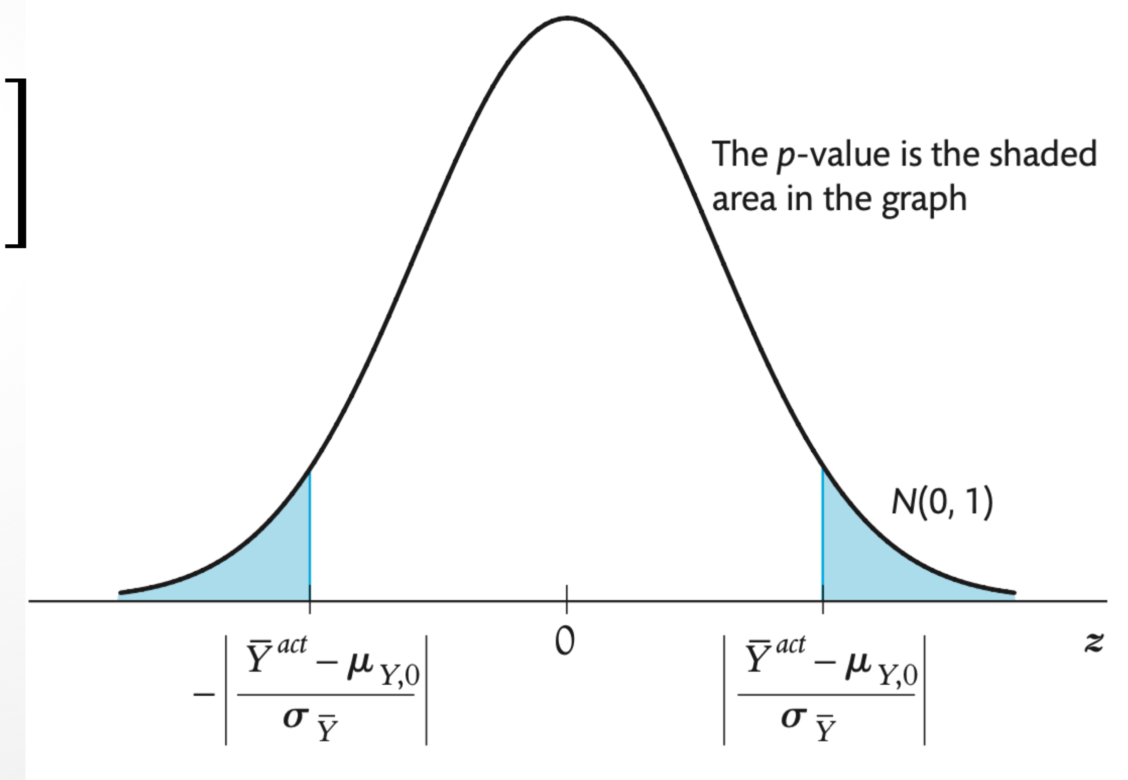
HOW TO CALCULATE THE P-VALUE

p-value = $Pr_{H_0}\left[\left|\bar{Y} - \mu_{Y,0}\right| > \left|\bar{Y}^{act} - \mu_{Y,0}\right|\right]$

$$= Pr_{H_0}\left[\left|\frac{\bar{Y}-\mu_{Y,0}}{\sigma_{\bar{Y}}}\right| > \left|\frac{\bar{Y}^{act}-\mu_{Y,0}}{\sigma_{\bar{Y}}}\right|\right]$$

$$= 2\Phi\left(-\left|\frac{\bar{Y}^{act}-\mu_{Y,0}}{\sigma_{\bar{Y}}}\right|\right)$$

- How to compute $\sigma_{\bar{Y}}$?
  - we know $\sigma_{\bar{Y}} = \frac{1}{\sqrt{n}}\sigma_Y \rightarrow$ we need σ_Y



SAMPLE VARIANCE

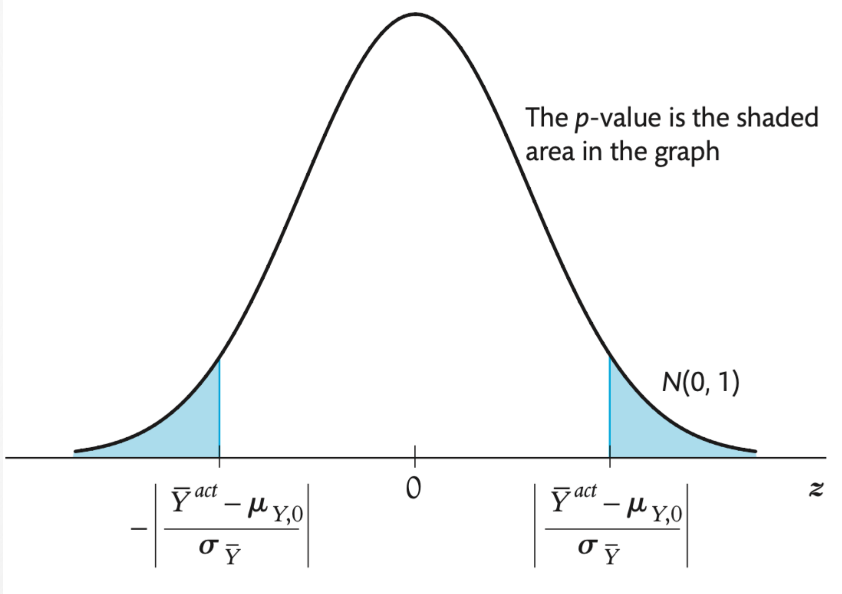
- $s_Y^2 = \frac{1}{n-1} \sum_{i=1}^{n} (Y_i - \bar{Y})^2$
- $E(s_Y^2) = \sigma_Y^2$
- $s_Y^2 \xrightarrow{p} \sigma_Y^2$

THE STANDARD ERROR OF $\bar{Y}$

- We need $\sigma_{\bar{Y}}$ to compute p-value.
- We know that $\sigma_{\bar{Y}} = \frac{1}{\sqrt{n}}\sigma_Y$
- We can estimate it using $\hat{\sigma} = \frac{1}{\sqrt{n}}s_Y$
- Called *standard error* of $\bar{Y}$: $SE(\bar{Y}) = \hat{\sigma} = \frac{1}{\sqrt{n}}s_Y$
- $SE(\bar{Y})$ measures the *precision* of $\bar{Y}$ as an estimate of μ_Y

HOW TO CALCULATE THE P-VALUE (2)

- p-value = $2\Phi\left(-\left|\frac{\bar{Y}^{act}-\mu_{Y,0}}{\sigma_{\bar{Y}}}\right|\right)$
- p-value= $2\Phi\left(-\left|\frac{\bar{Y}^{act}-\mu_{Y,0}}{SE(\bar{Y})}\right|\right) = 2\Phi(-|t|)$
- t $= \frac{\bar{Y}^{act}-\mu_{Y,0}}{SE(\bar{Y})}$ is the *t-statistic* (or *t-ratio*).



CALCULTING THE P-VALUE: AN EXAMPLE

- We have wages for a sample of 200 recent graduates
- $H_o: \mu_Y = \$20$
- In the sample, $\bar{Y}^{act} = \$22.64$; $s_Y = \$18.14$
- **YOUR TURN -** Calculate:
  1. $SE(\bar{Y})$,
  2. t-stat
  3. p-value

Remember:

- $SE(\bar{Y}) = \hat{\sigma} = \frac{1}{\sqrt{n}} s_Y$
- t-stat = $\frac{\bar{Y}^{act} - \mu_{Y,0}}{SE(\bar{Y})}$
- p-value = $2\Phi\left(-|t|\right)$

CALCULTING THE P-VALUE: AN EXAMPLE

- We have wages for a sample of 200 recent graduates
- $H_o: \mu_Y = \$20$
- In the sample, $\bar{Y}^{act} = \$22.64;\ s_Y = \$18.1$

- $SE(\bar{Y}) = \hat{\sigma} = \frac{1}{\sqrt{n}} s_Y = \frac{18.14}{\sqrt{200}} = 1.28$
- t-stat = $\frac{\bar{Y}^{act} - \mu_{Y,0}}{SE(\bar{Y})} = \frac{22.64-20}{1.28} = 2.06$
- p-value = $2\Phi\left(-|t|\right) = 2 * 0.0197 = 0.0394$

Accept or reject H_0?

SIGNIFICANCE LEVEL

- How low should the p-value be, for us to reject the null hypothesis?

- Convention in social sciences: 0.05 (or 5%)

$$\text{Reject } H_0 \text{if p} < 0.05 \ \rightarrow |t^{act}| > 1.96$$

- *5% significance level*
  - max probability of a *type-I error* we are willing to accept

3.3 CONFIDENCE INTERVALS

University *of* Massachusetts Amherst

CONFIDENCE INTERVALS

- **95% confidence interval:** a range of values that is 95% likely to include the population mean.

- The set of all values for μ_Y that we *cannot* reject at the 5% significance level.

- 95% confidence interval for μ_Y:

$$\bar{Y} - 1.96 * SE(\bar{Y}) \leq \mu_Y \leq \bar{Y} + 1.96 * SE(\bar{Y})$$

CONFIDENCE INTERVALS

YOUR TURN: Calculate a 95% confidence interval for hourly earnings

- In the sample, $\bar{Y}^{act} = \$22.64;\ \mathrm{SE}(\bar{Y}) = 1.28$

- *Reminder*: a 95% confidence interval for μ_Y is:

$$\bar{Y} - 1.96 * SE(\bar{Y}) \leq \mu_Y \leq \bar{Y} + 1.96 * SE(\bar{Y})$$

CONFIDENCE INTERVALS

YOUR TURN: Calculate a 95% confidence interval for hourly earnings

- In the sample, $\bar{Y}^{act} = \$22.64;\ \mathrm{SE}(\bar{Y}) = 1.28$
- Upper bound: $\bar{Y} + 1.96 * SE(\bar{Y})$ = 22.64 + 1.96$*$1.28 = 25.15
- Lower bound: $\bar{Y} - 1.96 * SE(\bar{Y})$ = 22.64 – 1.96$*$1.28 = 20.13
- $20.13 \leq \mu_Y \leq 25.15$

3.4 TESTING DIFFERENCES BETWEEN MEANS

University of Massachusetts Amherst

TESTING DIFFERENCES BETWEEN MEANS

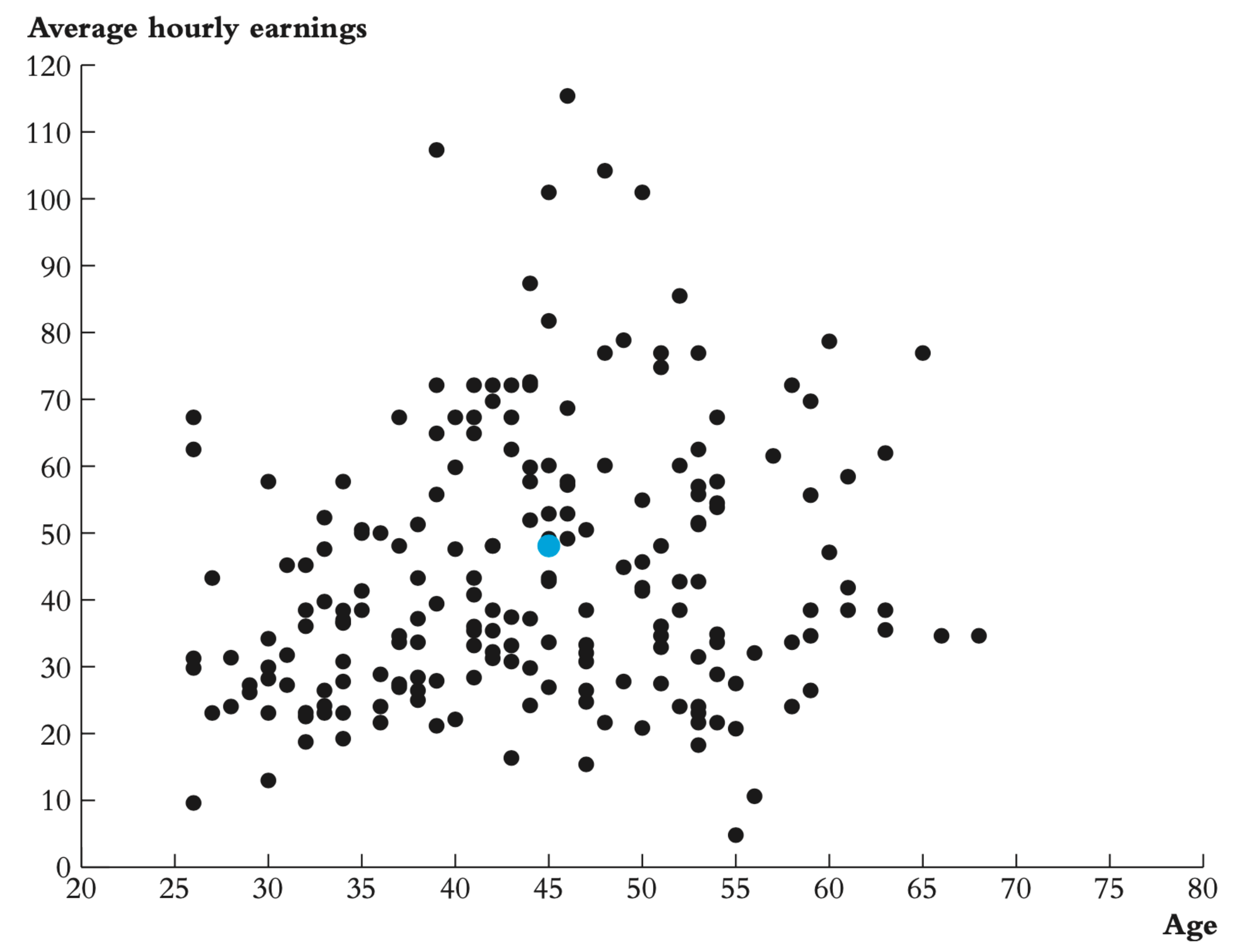
- $H_0\colon \mu_m - \mu_w = d_0$ vs. $H_1\colon \mu_m - \mu_w \neq d_0$
- $E(\bar{Y}_m - \bar{Y}_w) = \mu_m - \mu_w$
- $(\bar{Y}_m - \bar{Y}_w) \sim N\left(\mu_m - \mu_w, \frac{\sigma_m^2}{n_m} + \frac{\sigma_w^2}{n_w}\right)$
- $SE(\bar{Y}_m - \bar{Y}_w) = \sqrt{\frac{\sigma_m^2}{n_m} + \frac{\sigma_w^2}{n_w}}$
- $t = \frac{(\bar{Y}_m - \bar{Y}_w) - d_0}{SE(\bar{Y}_m - \bar{Y}_w)} \rightarrow p - value = 2\Phi(-|t^{act}|)$

3.5 SCATTERPLOTS AND SAMPLE CORRELATION

University *of* Massachusetts Amherst

SCATTERPLOTS

in STATA:
> scatter y x



SAMPLE COVARIANCE & CORRELATION

- (Population) Covariance & Correlation Coefficient:

$$cov(X,Y) = \sigma_{XY} = E[(X - \mu_X)(Y - \mu_Y)]$$

$$corr(X,Y) = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

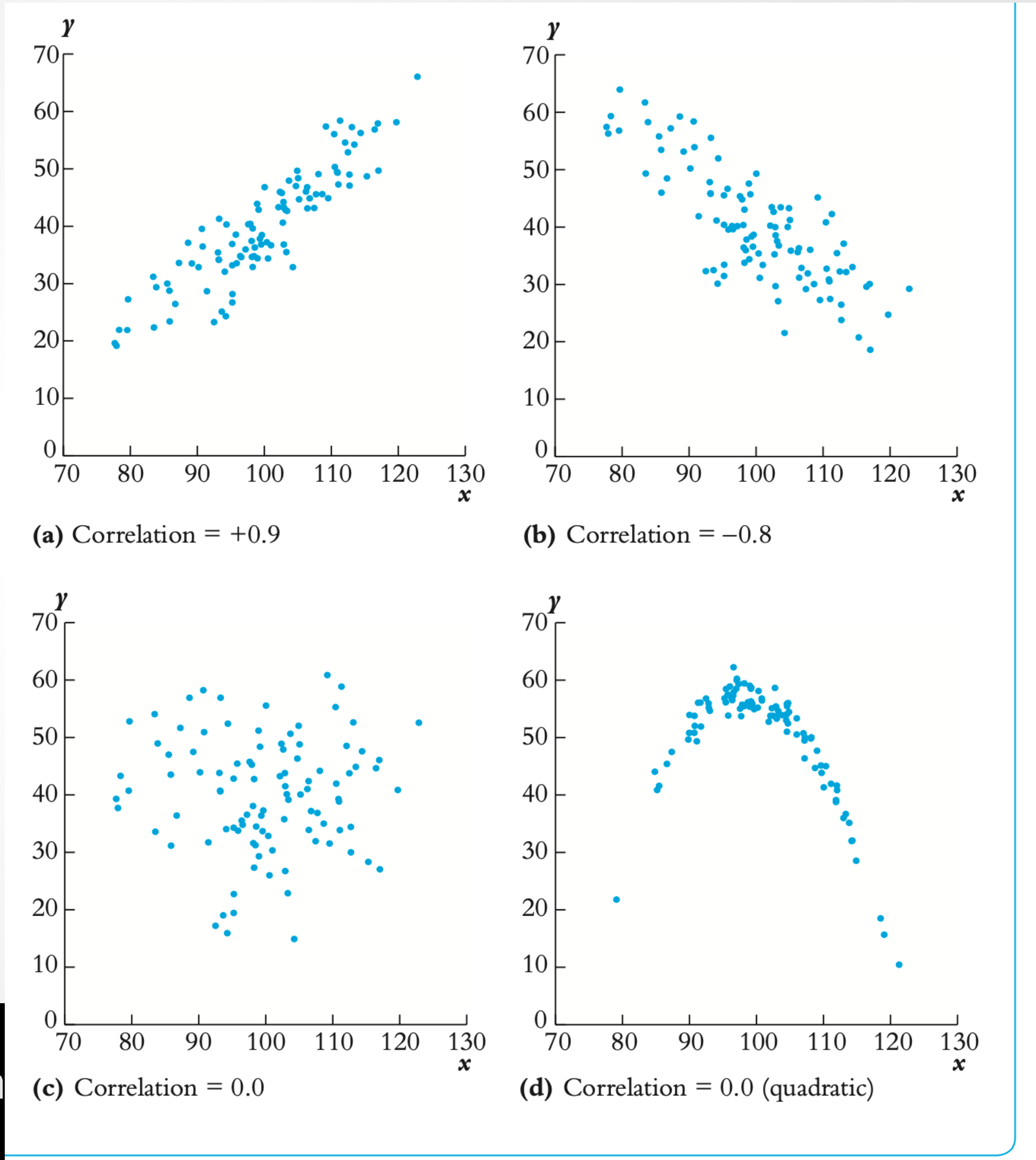
- Sample Covariance and Sample Correlation Coefficient:

$$s_{XY} = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})$$

$$r_{XY} = \frac{s_{XY}}{s_X s_Y}$$

SCATTERPLOTS & CORRELATION COEFFICIENTS

- The correlation coefficient captures *linear* associations between variables (as in panels (a) & (b)).
- It can miss non-linear ones (as in panel (d))



In the population of UMass students, the average number of study hours in the month of September is 100, with a variance of 43. In our usual notation, we can write $\mu_Y = 100$ and $\sigma_Y^2 = 43$.

If you take a random sample of 100 students and record their study hours in the month of September, what is the probability that the sample average is lower than 101? Formally, what is $Pr(\bar{Y} < 101)$?

(round up your answer to the 2nd decimal number)

$$\Pr(\bar{Y} < 101) = \Pr\left(\frac{\bar{Y} - \mu_Y}{\sigma_{\bar{Y}}} < \frac{101 - \mu_Y}{\sigma_{\bar{Y}}}\right)$$

$$\sigma_{\bar{Y}}^2 = \frac{\sigma_Y^2}{n} = \frac{43}{100} = 0.43$$

$$\Pr(\bar{Y} < 101) = \Pr\left(\frac{\bar{Y} - \mu_Y}{\sigma_{\bar{Y}}} < \frac{101 - 100}{\sqrt{0.43}}\right) = \Pr(z < 1.525) = \Phi(1.525) = 0.94$$