Today

- Confidence intervals
  - One proportion, one mean, two proportions, two means
- Confidence intervals and hypothesis tests
- Pairwise comparisons
- Correlation

Confidence intervals

Confidence intervals are random intervals that come with a long-run guarantee:

- *If you report 95% confidence intervals all your life, 95% of them will capture the true value*
- You can't say anything about a particular interval; it either contains the truth or it doesn't

Visualization: http://rpsychologist.com/d3/CI/

In SAS, you can find CIs for means and proportions (one and two groups) using the same PROCs we used for testing

One proportion

- PROC FREQ gives us intervals
- Example: drug.csv
- If we want 99% CI for recovery rate…

```
PROC FREQ data = drug;
TABLES recovery / binomial riskdiff alpha = 0.01;
RUN;
```

alpha is "1- confidence level" [here conf. level = 0.99]

Binomial Proportion	
recovery = 0	
Proportion	0.5667
ASE	0.0640
99% Lower Conf Limit	0.4019
99% Upper Conf Limit	0.7315
Exact Conf Limits	
99% Lower Conf Limit	0.3938
99% Upper Conf Limit	0.7287

Based on normal approximation (you probably saw this one in intro stats)

Doesn't rely on normal approximation

One mean

- PROC TTEST gives us CIs for means
- Example: speed dataset
- 95% confidence interval for max. speed

```
PROC TTEST data = speed alpha = 0.05;
    VAR speed;
RUN;
```

Mean	95% CL Mean		Std Dev	95% CL Std Dev	
90.7330	89.5166	91.9493	22.4157	21.5882	23.3098

Two independent means

- Again, use PROC TTEST
- Example:
  - 99% CI for difference in max speed "female – male"

```
PROC TTEST data = speed alpha = 0.01;
     VAR speed;
     CLASS gender;
RUN;
```

gender	Method	Mean	99% CL Mean		Std Dev	99% CL Std Dev	
female		87.0865	85.2087	88.9643	21.4179	20.1650	22.8244
male		97.9182	95.1278	100.7	22.6250	20.8068	24.7641
Diff (1-2)	Pooled	-10.8317	-14.1280	-7.5353	21.8314	20.7804	22.9863
Diff (1-2)	Satterthwaite	-10.8317	-14.1903	-7.4730			

Equal variance (Pooled)

Unequal variance (Satterthwaite)

Two proportions

- Use PROC FREQ
- Example: 2drugs.csv
- 99% CI for difference in recovery rates

```
PROC FREQ data = twodrugs;
TABLES recovery*drug / chisq riskdiff alpha = 0.01;
RUN;
```

Column 1 Risk Estimates						
	Risk	ASE	(Asymptotic) 99% Confidence Limits		(Exact) 99% Confidence Limits	
Row 1	0.3571	0.0906	0.1239	0.5904	0.1477	0.6155
Row 2	0.5556	0.0956	0.3092	0.8019	0.3002	0.7912
Total	0.4545	0.0671	0.2816	0.6275	0.2831	0.6340
Difference	-0.1984	0.1317	-0.5376	0.1408		
Difference is (Row 1 - Row 2)						

CIs and hypothesis tests

- ***Example:*** Want to know if the difference in math scores between men and women is significantly different than 0 at the 0.05
  - ***We can find a 95% confidence interval for the difference in scores "men – women" and check whether it contains 0***
  - If the interval contains 0, don't reject the null hypothesis that there is no difference
  - If the interval doesn't contain 0, there are significant differences between men and women at the 0.05 significance level

CIs and hypothesis tests

In general…

- Let θ be an unknown feature of the DGM
- Suppose we know how to construct (1-α)100% confidence intervals for θ
- We want to test H_0: $\theta = \theta_0$ against Ha: $\theta \neq \theta_0$ at the α significance level
- We can do the test by checking whether θ_0 is contained in the interval
- If θ_0 is in the interval, don't reject the null; otherwise, reject the null

Pairwise comparisons

Comparing more than 2 groups

- ***Example:*** We want to know if there are differences in average standardized testing scores for different socioeconomic statuses, using the hsb2 dataset
- How can we solve this problem?
- We know how to compare 2 groups, but now we have 3 groups: low, middle, and high socioeconomic status

Pairwise tests

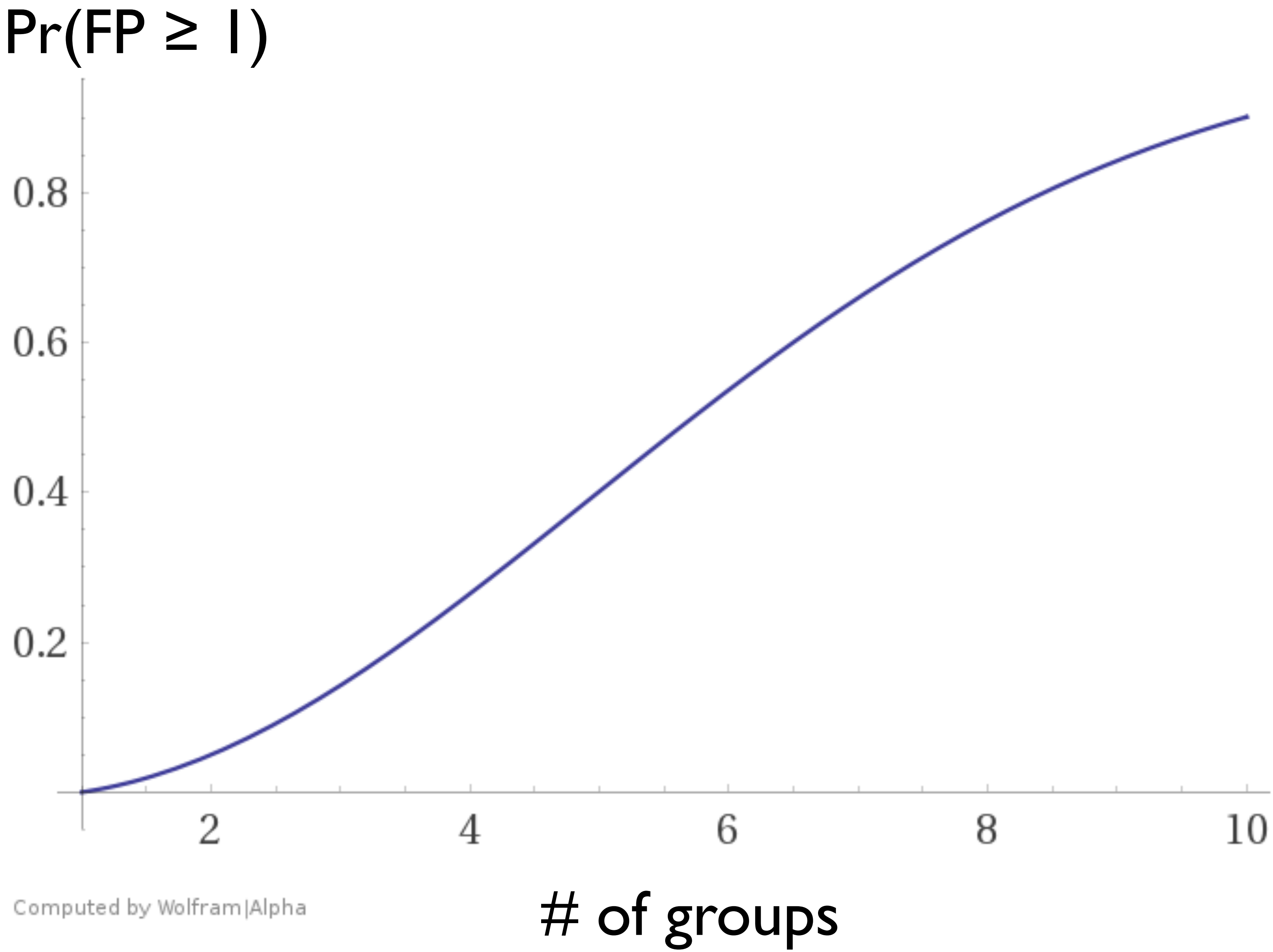
- An approach is doing 3 pairwise two-sample tests
  - Low vs middle
  - Middle vs high
  - Low vs high
- If we do these 3 tests at the 0.05 significance level (each), the probability that there is at least one false positive (type I error) is roughly 0.14

Pairwise tests

- If we have *k* groups, there are *k* choose 2 pairwise comparisons
- If our significance level is 0.05, the probability that there's at least one false positive (FP) is

$$\Pr(FP \geq 1) = 1 - \Pr(FP = 0) = 1 - 0.95^{(k \text{ choose } 2)}$$

- For example, if k = 5, Pr(FP ≥ 1) is approximately 0.4


Pr(FP ≥ 1)
0.8
0.6
0.4
0.2
2
4
6
8
10
Computed by Wolfram|Alpha
# of groups

A (not-so-great) fix

- A general solution to this "multiple testing" problem (which isn't specific for pairwise comparisons) is the following
- **Bonferroni**: If we're are doing N tests and want to ensure an overall false positive rate of 0.05, conduct the individual tests at the 0.05/N significance level
- **Problem**: Very stringent.
  - For example, if we have 5 groups, there are N = (5 choose 2) = 10 pairwise tests, so we should perform the tests at the 0.005 significance level, which is quite harsh

Tukey’s honest significant difference

- If we’re comparing the “means” (expectations) of groups with either / or
  - Approximately normal distributions
  - Sample sizes that are big enough, and the DGM has finite variance
- We can use a less stringent method called Tukey’s honest significant difference (there are others)
- SAS will do it for us

- ***Example:*** compare average scores in standardized tests for low, middle and high socioeconomic status at an overall significance level α = 0.01

```
PROC ANOVA data = hsbnew;
    CLASS ses;
    MODEL avg = ses;
    MEANS ses / Tukey alpha = 0.01;
RUN;
```

Alpha	0.01
Error Degrees of Freedom	197
Error Mean Square	59.57878
Critical Value of Studentized Range	4.16833

Comparisons significant at the 0.01 level are indicated by ***.

ses Comparison	Difference Between Means	Simultaneous 99% Confidence Limits		
high - middle	4.344	0.553	8.135	***
high - low	7.617	3.152	12.082	***
middle - high	-4.344	-8.135	-0.553	***
middle - low	3.274	-0.784	7.331	
low - high	-7.617	-12.082	-3.152	***
low - middle	-3.274	-7.331	0.784	

If an interval doesn't contain 0, the difference between the group is significant

Correlation

Sample correlation

- Sample correlation is useful for quantifying the degree of ***linear*** association between 2 quantitative variables
- It can be computed in different equivalent ways. For example, if we have variables *X* and *Y* that come in pairs:

$$(x_1, y_1), (x_2, y_2), \ \ldots, (x_n, y_n)$$

- We can compute a z-score for each datum:

$$z_{x_i} = \frac{x_i - \overline{x}}{s_x} \qquad z_{y_i} = \frac{y_i - \overline{y}}{s_y}$$

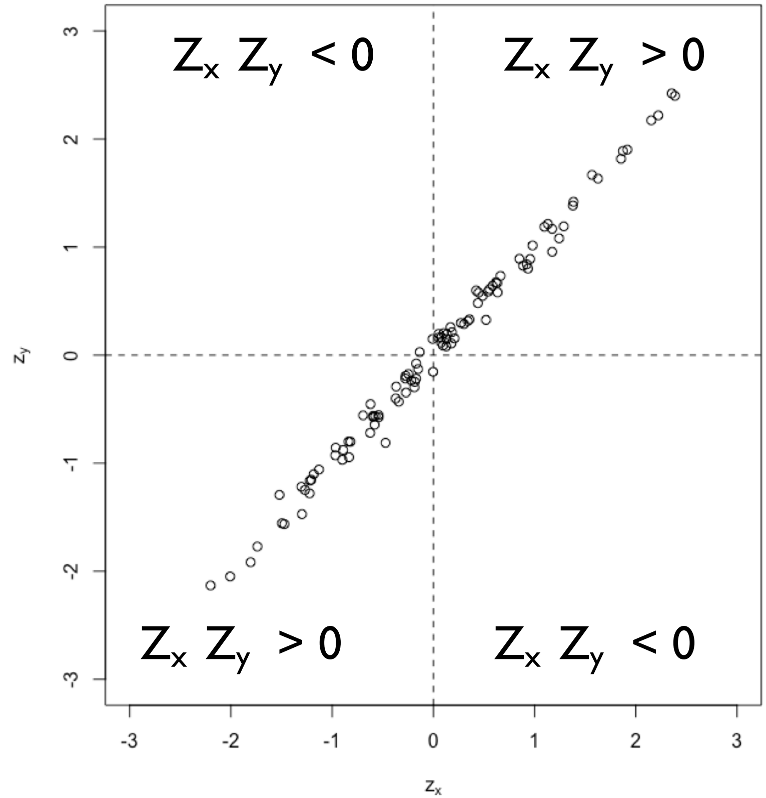
- And find:

$$r = \frac{1}{n-1}(z_{x_1}z_{y_1} + z_{x_2}z_{y_2} + \cdots + z_{x_n}z_{y_n})$$

Correlation

- *r* is always between -1 and 1. The extremes are attained when there are perfect linear relationships (with negative and positive slope, respectively)

Positive correlation ($r > 0$)



When x_i is above the mean of x, y is usually above the mean of y_i

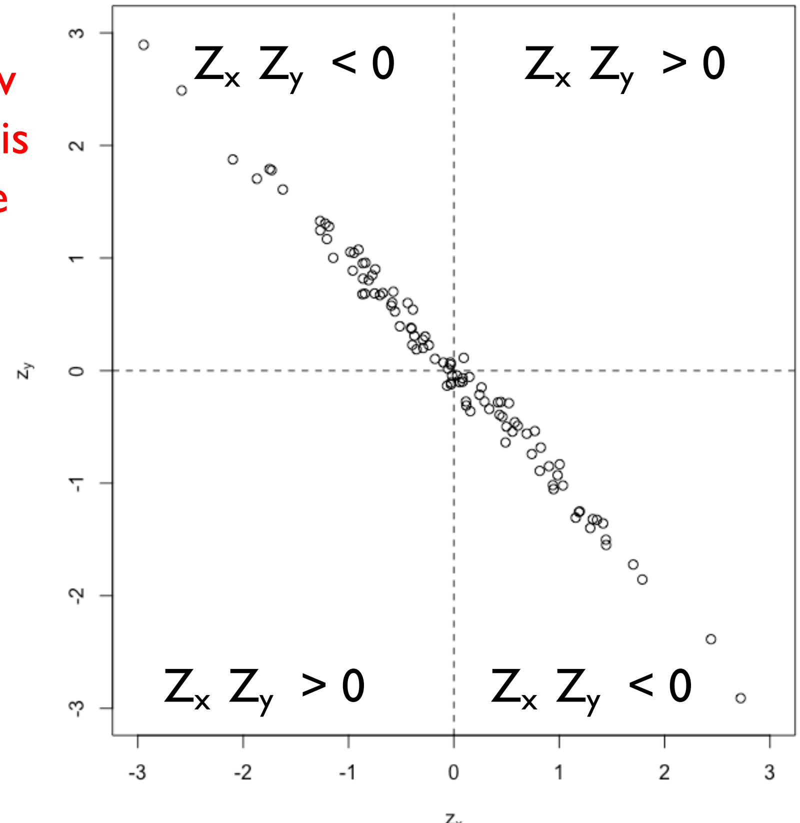
When x_i is below the mean of x, y is usually below the mean of y_i

mostly positive

$$r = \frac{1}{n-1}\left(z_{x_1}z_{y_1} + z_{x_2}z_{y_2} + \cdots + z_{x_n}z_{y_n}\right)$$

Negative correlation *(r < 0)*

When x_i is below the mean of x, y is usually above the mean of y_i

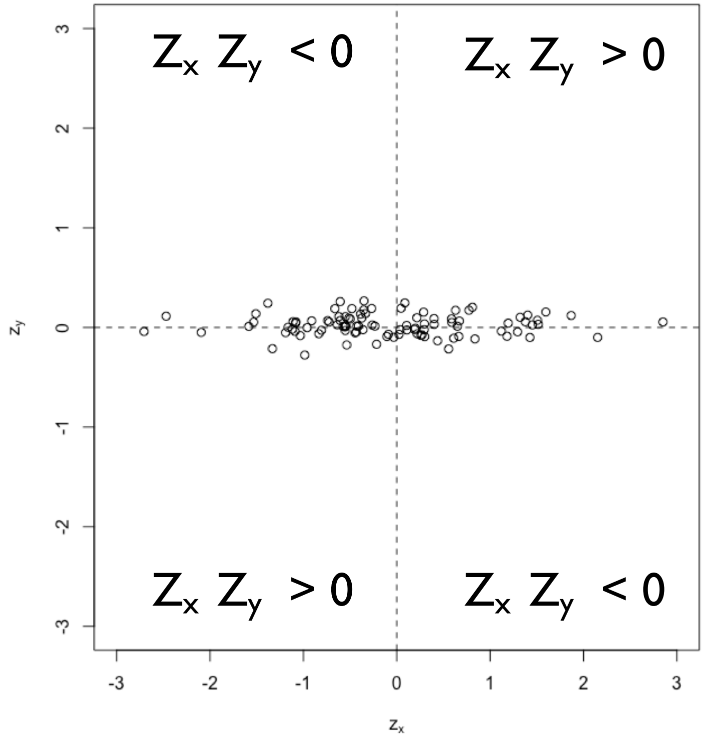


When x_i is above the mean of x, y is usually below the mean of y_i

mostly negative

$$r = \frac{1}{n-1}\underbrace{(z_{x_1}z_{y_1} + z_{x_2}z_{y_2} + \cdots + z_{x_n}z_{y_n})}$$

Correlation ~ 0

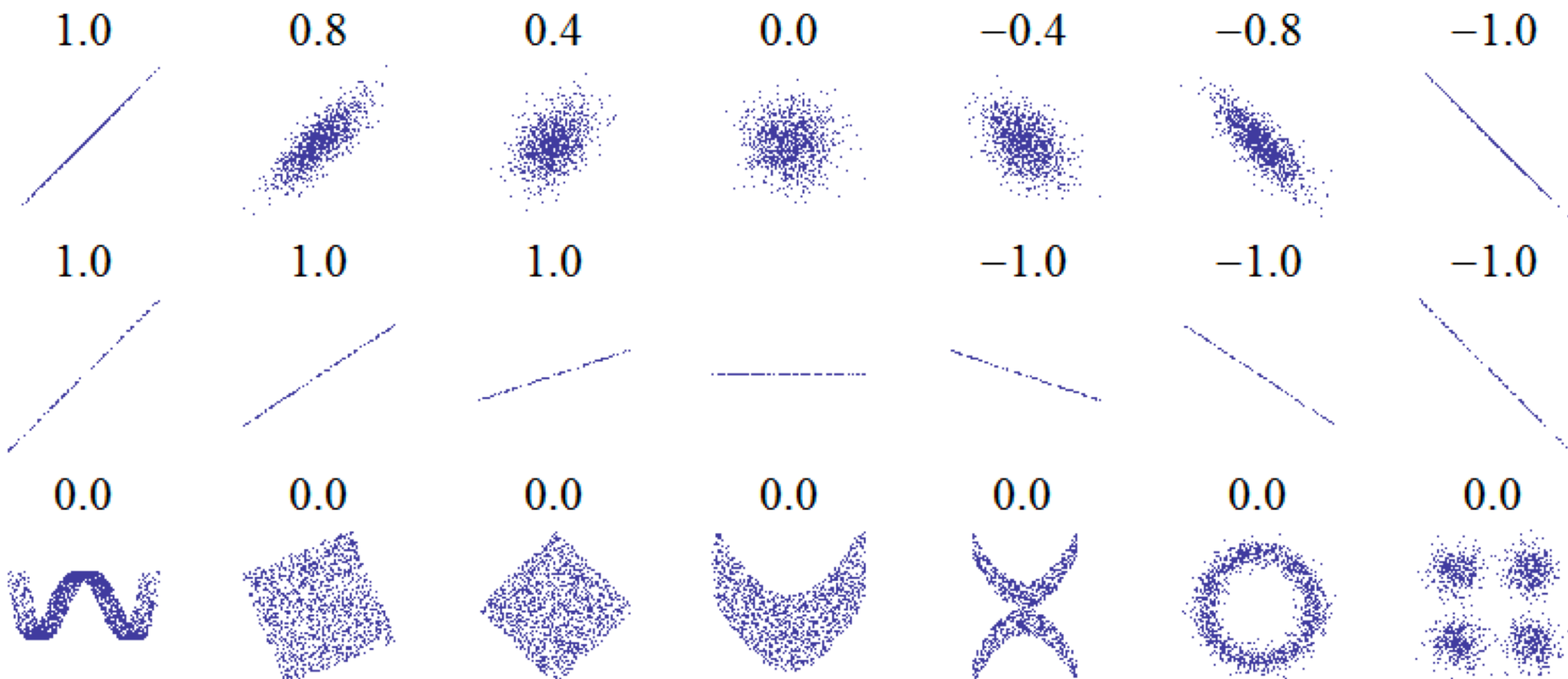


roughly the same positive & negative… will cancel out & r ~ 0

$$r = \frac{1}{n-1}\left(z_{x_1}z_{y_1} + z_{x_2}z_{y_2} + \cdots + z_{x_n}z_{y_n}\right)$$

r measures the strength and direction of **linear dependence**:

- *If there is a clear pattern, but it isn't linear… r is inadequate!*



https://en.wikipedia.org/wiki/Correlation_and_dependence

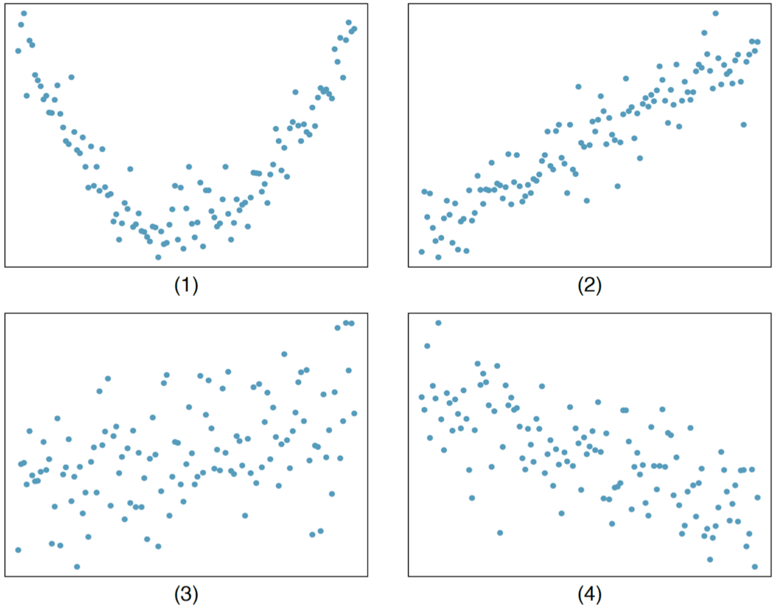
7.7 Match the correlation, Part I. Match the calculated correlations to the corresponding scatterplot.

(a) $r = -0.7$

(b) $r = 0.45$

(c) $r = 0.06$

(d) $r = 0.92$



https://www.openintro.org/

Correlations with SAS

- PROC CORR computes correlations for us
- To visualize the data, we can create a "scatterplot matrix" with PROC SGSCATTER
- ***Example:*** in the hsb2 dataset, suppose that we want to find the pairwise correlations between math, writing, reading, science, and social studies scores

```
PROC CORR data = hsbnew;
VAR math write socst science read;
RUN;

PROC SGSCATTER data = hsbnew;
matrix math write socst science read;
RUN;
```

Pearson Correlation Coefficients, N = 200 Prob > |r| under H0: Rho=0					
	math	**write**	**socst**	**science**	**read**
math	1.00000	0.61745	0.54448	0.63073	0.66228
		<.0001	<.0001	<.0001	<.0001
write	0.61745	1.00000	0.60479	0.57044	0.59678
	<.0001		<.0001	<.0001	<.0001
socst	0.54448	0.60479	1.00000	0.46511	0.62148
	<.0001	<.0001		<.0001	<.0001
science	0.63073	0.57044	0.46511	1.00000	0.63016
	<.0001	<.0001	<.0001		<.0001
read	0.66228	0.59678	0.62148	0.63016	1.00000
	<.0001	<.0001	<.0001	<.0001