Sections 2.5 and 2.6

Note made by: Dr. Timothy Hanson
Instructor: Peijie Hou

Department of Statistics, University of South Carolina

Stat 205: Elementary Statistics for the Biological and Life Sciences

2.5 Relationships between variables

- Histogram, boxplot, mean, & five-number summary all summarize one numeric variable.
- Bar chart & relative frequency tables summarize one categorical variable.
- How about exploring the relationship between two (or more) variables?
- *Usually* more than one variable is measured on each observational unit because we want to quantify and test *associations* between two (or more) variables.

Categorical-categorical

- When we have two categorical variables we can *cross-classify* them in a **contingency table.**
- List categories of one variable along top, categories of other along side, and count number falling into each *pair* of categories.
- Contingency tables will come back in Chapters 3 and 10.

Example 2.5.1: *E. Coli* watershed contamination

- $n = 623$ water specimens collected at three locations that feed into Morro Bay (north of L.A.): Chorro Creek ($n_1 = 241$), Los Osos Creek ($n_2 = 256$), and Baywood Seeps ($n_3 = 126$). Using DNA fingerprinting, the origin of the *E. Coli* was found: bird, cat or dog, farm animal, humans, other wild mammals. (p. 53).
- Two variables: **creek** and **source**. The data are naturally cross-classified into a contingency table:

Table 2.5.1 Frequency table of *E. coli* source by location

	E. Coli Source					
Location	Bird	Domestic pet	Farm animal	Human	Terrestrial mammal	Total
Chorro Creek	46	29	106	38	22	**241**
Los Osos Creek	79	56	32	63	26	**256**
Baywood	35	23	0	60	8	**126**
Total	**160**	**108**	**138**	**161**	**56**	**623**

Relative frequencies within Creek

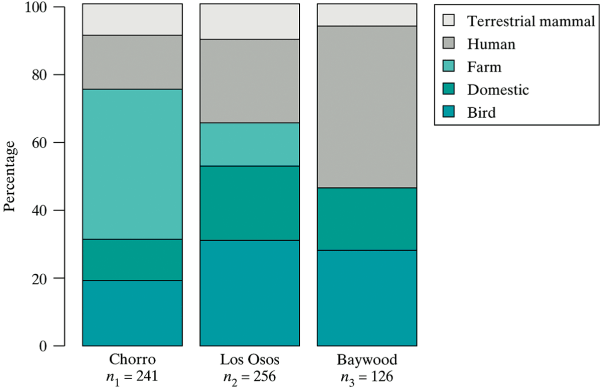
- Relative frequencies highlight differences in *E. Coli* sources across the Creeks.
- Are domestic pets more of an *E. Coli* problem (o.e., source) at Chorro Creek or Baywood?

Table 2.5.2 Bivariate relative frequency table (row percentages) of *E. coli* source by location

	E. Coli Source					
Location	Bird	Domestic pet	Farm animal	Human	Terrestrial mammal	Total
Chorro Creek	19.1	12.0	44.0	15.8	9.1	**100**
Los Osos Creek	30.9	21.9	12.5	24.6	10.2	**100**
Baywood	27.8	18.3	0.0	47.6	6.3	**100**
Total	**25.7**	**17.3**	**22.2**	**25.8**	**9.0**	**100**

Stacked relative frequency bar chart

- All %'s of *E. Coli* source add to 100% within a creek.
- Bar chart showing relative proportions within each creek lets us visually compare, e.g., which creek has a higher percentage of farm waste (Chorro).



Discussion

- Visually, the stacked relative bar chart shows differences in *E. coli* source across creeks. Differences may be real, or due to sampling variability.
- A scientist would like to test the hypothesis that the three streams have different percentages of *E. Coli* sources.
- This is done with a chi-squared test (χ^2 test); covered in Chapter 10.
- The real question being asked is whether there is a verifiable *association* between creek and origins of *E. Coli* present.

Categorical-numeric relationships

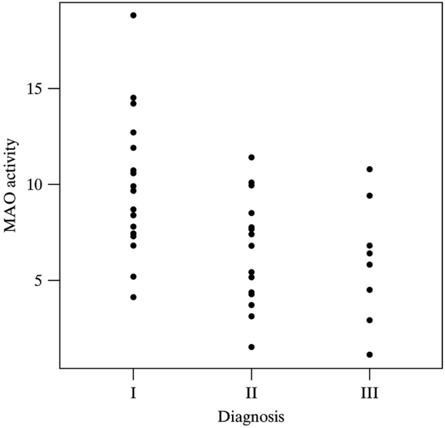


Figure 1.1.2 MAO activity in schizophrenic patients

- Example 1.1.4: MAO and schizophrenia.
- Side-by-side dotplots.

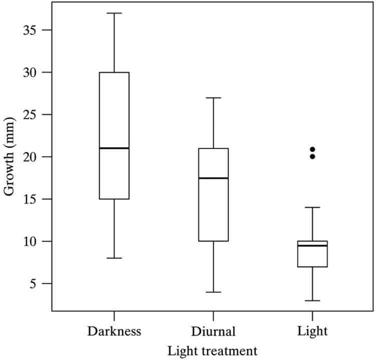
Categorical-numeric relationships

- Previous slide: MAO activity (continuous numeric) and schizophrenia diagnosis (ordinal categorical): I, II, III.
- Typically want to describe how one variable is changing with the other.
- For example, activity seems to decrease from I to II to III.
- Put another way, the *probability* of the more severe categories II and III increases as activity decreases (need to ponder that one a bit).

Formal approaches for numeric-categorical

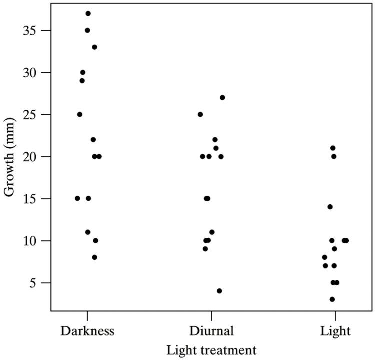
- Modeling activity as a function of schizophrenia category is done with analysis of variance in Chapter 11.
- Modeling the probability of severe diagnosis as a function of activity is done with logistic regression in Chapter 12.
- Either way we are again looking for an *association* between MAO activity and diagnosis I, II, III.
- R code for dotplot in Lecture 1 slides. Boxplots from `boxplot(mao~group)`.

Example 2.5.3 Radish growth



- Lengths of radish shoots grown in three lighting situations: total darkness, half-and-half, total light.
- How is shoot growth related to light? Is the observed pattern due to chance alone, or does light alter initial growth?

Radish growth, continued



Side-by-side dotplots give similar information.

Numeric-numeric relationships

- When both variables are continuous, the underlying relationship may be smooth.
- A scatterplot shows each pair of numeric-variables across experimental units.
- If the scatterplot shows an obvious pattern, the pattern can be *modeled* by a function, often just a line.
- Scatterplots rarely show a perfect relationship, but rather some sort of smooth association plus noise.
- It's that "plus noise" where statistic come in. We need to separate the *signal* from the *noise*.

Example 2.5.4 Whale selenium

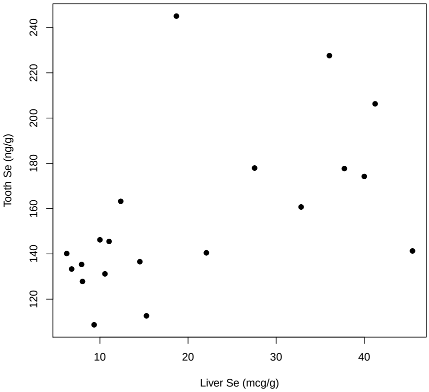
- Selenium protects marine animals against mercury poisoning.
- $n = 20$ Beluga whales were sampled during an Eskimo hunt; tooth Selenium (Se) and liver Se were measured.
- Are tooth Se and liver Se are related? How? Focus of Chapter 12.

Table 2.5.3 Liver and tooth selenium concentrations of twenty belugas

Whale	Liver Se (μg/g)	Tooth Se (ng/g)	Whale	Liver Se (μg/g)	Tooth Se (ng/g)
1	6.23	140.16	11	15.28	112.63
2	6.79	133.32	12	18.68	245.07
3	7.92	135.34	13	22.08	140.48
4	8.02	127.82	14	27.55	177.93
5	9.34	108.67	15	32.83	160.73
6	10.00	146.22	16	36.04	227.60
7	10.57	131.18	17	37.74	177.69
8	11.04	145.51	18	40.00	174.23
9	12.36	163.24	19	41.23	206.30
10	14.53	136.55	20	45.47	141.31

Example 2.5.4 R code

```
liver=c(  6.23,  6.79,  7.92,  8.02,  9.34, 10.00, 10.57, 11.04, 12.36, 14.53,
         15.28, 18.68, 22.08, 27.55, 32.83, 36.04, 37.74, 40.00, 41.23, 45.47)
tooth=c(140.16,133.32,135.34,127.82,108.67,146.22,131.18,145.51,163.24,136.55,
        112.63,245.07,140.48,177.93,160.73,227.60,177.69,174.23,206.30,141.31)
plot(liver,tooth,xlab="Liver Se (mcg/g)",ylab="Tooth Se (ng/g)",pch=19)
```



2.6 Measures of dispersion

- The mean and median measure central tendancy, i.e. give "typical" values.
- Also need to know how spread out or variable the data are. This gives information on how "close" data values are to "typical."
- The sample range is the length of the interval that contains all of the data range $=$ max $-$ min.
- The range is sensitive to very large or small values; the IQR is less sensitive, or *more robust* to outlying values.
- The **sample variance** is the average squared deviation

$$s^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}.$$

Sample standard deviation

The **sample standard deviation** is the square root of the variance

$$s = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}}.$$

- It is a pain to compute by hand.
- Need to (1) find mean, (2) subtract mean off each value, (3) square the deviations about the mean, (4) add the squared deviations up, (5) divide by $n - 1$, then (6) take the square root.

Example 2.6.4 Chrysanthemum growth

A botanist measured stem elongation (mm) over a week of five plants on the same greenhouse bench

$$76 \quad 72 \quad 65 \quad 70 \quad 82.$$

The sample mean is

$$\bar{y} = \frac{76 + 72 + 65 + 70 + 82}{5} = \frac{365}{5} = 73\text{mm}.$$

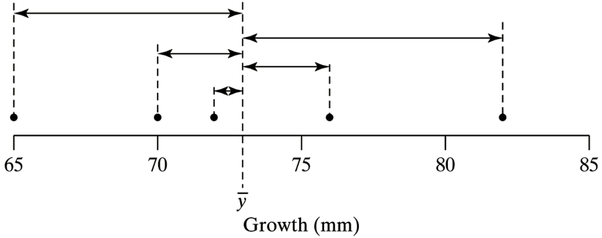
Computing the standard deviation

Table 2.6.1 Illustration of the formula for the sample standard deviation

Observation (y_i)	Deviation ($y_i - \bar{y}$)	Squared deviation $(y_i - \bar{y})^2$
76	3	9
72	−1	1
65	−8	64
70	−3	9
82	9	81
Sum $365 = \sum_{i=1}^{n} y_i$	0	$164 = \sum_{i=1}^{n} (y_i - \bar{y})^2$

$$s = \sqrt{\frac{164}{4}} = \sqrt{41} = 6.4 \text{ mm}.$$

Chrysanthemum SD



- The standard deviation (SD) is like the average deviation around the mean (but not quite).
- Need to *square* the $n = 5$ lengths in plot, take average, then square root.
- The SD and the IQR have same units as data.
- Measures how spread out observations typically are.

Empirical rule

Typical Percentages: The Empirical Rule

For "nicely shaped" distributions—that is, unimodal distributions that are not too skewed and whose tails are not overly long or short—we usually expect to find

about 68% of the observations within ± 1 SD of the mean.

about 95% of the observations within ± 2 SDs of the mean.

$>$99% of the observations within ± 3 SDs of the mean.

Example 2.6.8 Weight gain for bulls

- Average weight gain over 140 days for $n = 39$ Charolais bulls recorded.
- Five number summary: 1.18, 1.29, 1.41, 1.58, 1.92 kg/day.

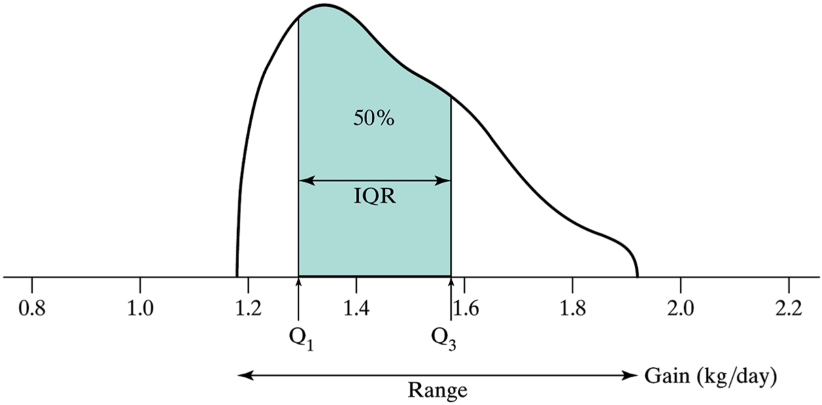
Table 2.6.2 Average daily gain (kg/day) of thirty-nine Charolais bulls

1.18	1.24	1.29	1.37	1.41	1.51	1.58	1.72
1.20	1.26	1.33	1.37	1.41	1.53	1.59	1.76
1.23	1.27	1.34	1.38	1.44	1.55	1.64	1.83
1.23	1.29	1.36	1.40	1.48	1.57	1.64	1.92
1.23	1.29	1.36	1.41	1.50	1.58	1.65	

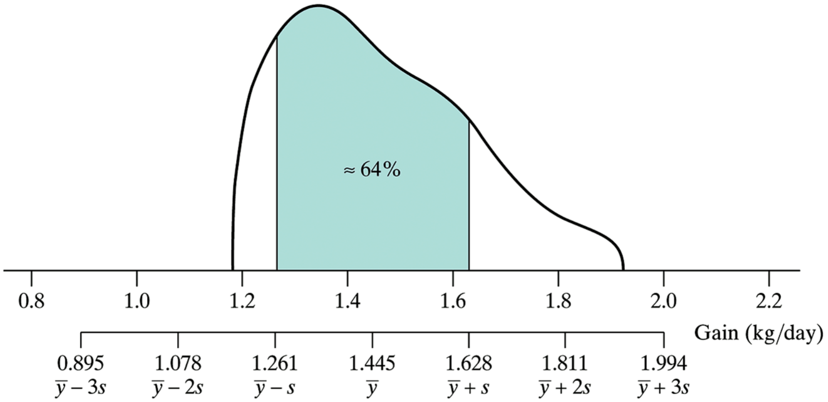
Bull weight gain R code

```
> bulls=c(1.18,1.24,1.29,1.37,1.41,1.51,1.58,1.72,
+         1.20,1.26,1.33,1.37,1.41,1.53,1.59,1.76,
+         1.23,1.27,1.34,1.38,1.44,1.55,1.64,1.83,
+         1.23,1.29,1.36,1.40,1.48,1.57,1.64,1.92,
+         1.23,1.29,1.36,1.41,1.50,1.58,1.65)
> mean(bulls)
[1] 1.444615
> sd(bulls)
[1] 0.1831285
> median(bulls)
[1] 1.41
> IQR(bulls)
[1] 0.285
> range(bulls)
[1] 1.18 1.92
> max(bulls)-min(bulls)
[1] 0.74
```

Bulls: dispersion based on five-number summary



Bulls: dispersion based on mean and SD



Discussion

- Range easiest to understand, but largely driven by outliers. Only looks at two data points!
- Interquartile range also easy to understand, gives length of middle 50% of data.
- SD uses all data but also prone to inflation by outliers.
- However, SD and mean are "classical" measures and form basis of computing confidence intervals, and hypothesis tests, so we'll mainly use these from now on.