Stat 13, Intro. to Statistical Methods for the Life and Health Sciences.

1. Return midterm.
2. Comparing 2 props. with theory-based testing, smoking and gender example.
3. Five number summary, IQR, and geysers.
4. t-test, breastfeeding and intelligence.

No lecture Tue May 12!!!

Read ch7 and 10.

HW2, due Fri, May8, 1159pm. 2.3.15, 3.3.18, and 4.1.23.

The course website is http://www.stat.ucla.edu/~frederic/13/S26 .

Smoking and Gender

- Fukuda et al. (2002) found the following in Japan.
  - Out of 3602 births where both parents did not smoke, 1975 were boys. This 54.8% boys.
  - Out of 565 births where both parents smoked more than a pack a day, 255 were boys. This is 45.1% boys.
  - In total, out of 4170 births, 2230 were boys, which is 53.5% boys.

Formulas

- How do we find the margin of error for the difference in proportions?

$$Multiplier \times \sqrt{\left(\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}\right)}$$

- The multiplier is from the normal distribution and is dependent upon the confidence level.
  - 1.645 for 90% confidence
  - 1.96 for 95% confidence
  - 2.576 for 99% confidence

- We can write the confidence interval in the form:
  - statistic ± margin of error.

Smoking and Gender

- Our statistic is the observed sample difference in proportions, 0.097.
  - Plugging in $1.96 \times \sqrt{\left(\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}\right)} = 0.044$, we get 0.097 ± 0.044 as our 95% CI.
  - We could also write this interval as (0.053, 0.141).
  - We are 95% confident that the probability of a boy baby where neither family smokes minus the probability of a boy baby where both parents smoke is between 0.053 and 0.141.

A clarification on the formulas

- For CIs, the margin of error for the difference in proportions is $Multiplier \times \text{SE}$, where $\text{SE} = \sqrt{\left(\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}\right)}$

  In testing, the null hypothesis is no difference between the two groups, so we use the SE

  $$\sqrt{\left(\frac{\hat{p}(1-\hat{p})}{n_1} + \frac{\hat{p}(1-\hat{p})}{n_2}\right)}$$

  where $\hat{p}$ is the proportion in both groups combined. But in CIs, we use the formula $\sqrt{\left(\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}\right)}$ because we are not assuming $\hat{p}_1 = \hat{p}_2$ with CIs.

Smoking and Gender

- How would the interval change if the confidence level was 99%?
- The SE = $\sqrt{\left(\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}\right)}$ = .0224.
- Previously, for a 95% CI, it was 0.097 ± 1.96 x .0224 = 0.097 ± 0.044.
- For a 99% CI, it is 0.097 ± 2.576 x .0224 = 0.097 ± 0.058.

Smoking and Gender

- Written as the statistic ± margin of error, the 99% CI for the difference between the two proportions is

$$0.097 \pm 0.058.$$

- Margin of error
  - 0.058 for the 99% confidence interval
  - 0.044 for the 95% confidence interval

Smoking and Gender

- How would the 95% confidence interval change if we were estimating

$$\pi_{\text{smoker}} - \pi_{\text{nonsmoker}}$$

instead of

$$\pi_{\text{nonsmoker}} - \pi_{\text{smoker}} \text{ ?}$$

Smoking and Gender

- (−0.141, −0.053) or -0.097 ± 0.044

  instead of
- (0.053, 0.141) or 0.097 ± 0.044.

- The negative signs indicate the probability of a boy born to smoking parents is lower than that for nonsmoking parents.

Smoking and Gender

Validity Conditions of Theory-Based

- Same as with a single proportion.
- Should have at least 10 observations in each of the cells of the 2 x 2 table.

	Smoking Parents	Non-smoking Parents	Total
Male	255	1975	2230
Female	310	1627	1937
Total	565	3602	4167

Smoking and Gender

- The strong significant result in this study yielded quite a bit of press when it came out.
- Soon other studies came out which found no relationship between smoking and gender (Parazinni et al. 2004, Obel et al. 2003).
- James (2004) argued that confounding variables like social factors, diet, environmental exposure or stress were the reason for the association between smoking and gender of the baby. These are all confounded since it was an observational study. Different studies could easily have had different levels of these confounding factors.

Five number summary, IQR, and geysers.

6.1: Comparing Two Groups: Quantitative Response
6.2: Comparing Two Means: Simulation-Based Approach
6.3: Comparing Two Means: Theory-Based Approach

Exploring Quantitative Data

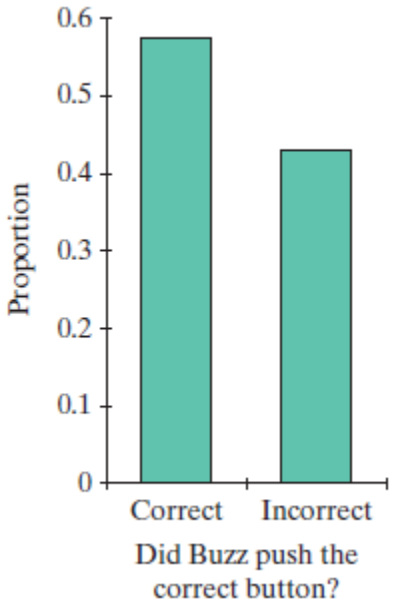
Section 6.1

Quantitative vs. Categorical Variables

- Categorical
  - Values for which arithmetic does not make sense.
  - Gender, ethnicity, eye color…
- Quantitative
  - You can add or subtract the values, etc.
  - Age, height, weight, distance, time…

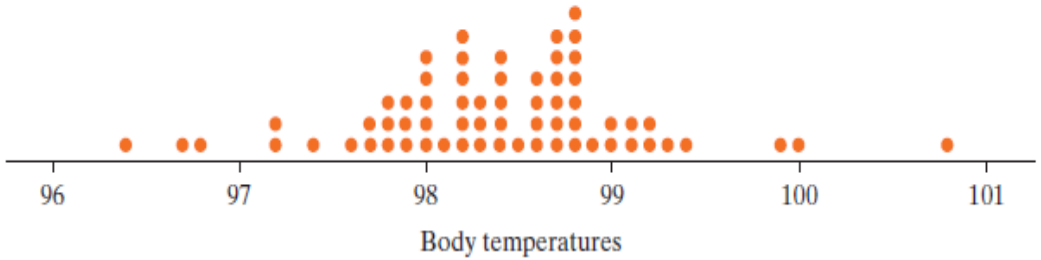
Graphs for a Single Variable

Categorical



Bar Graph

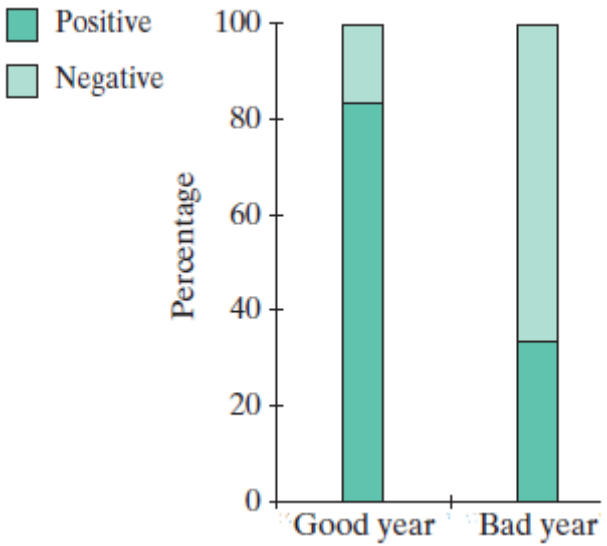
Quantitative



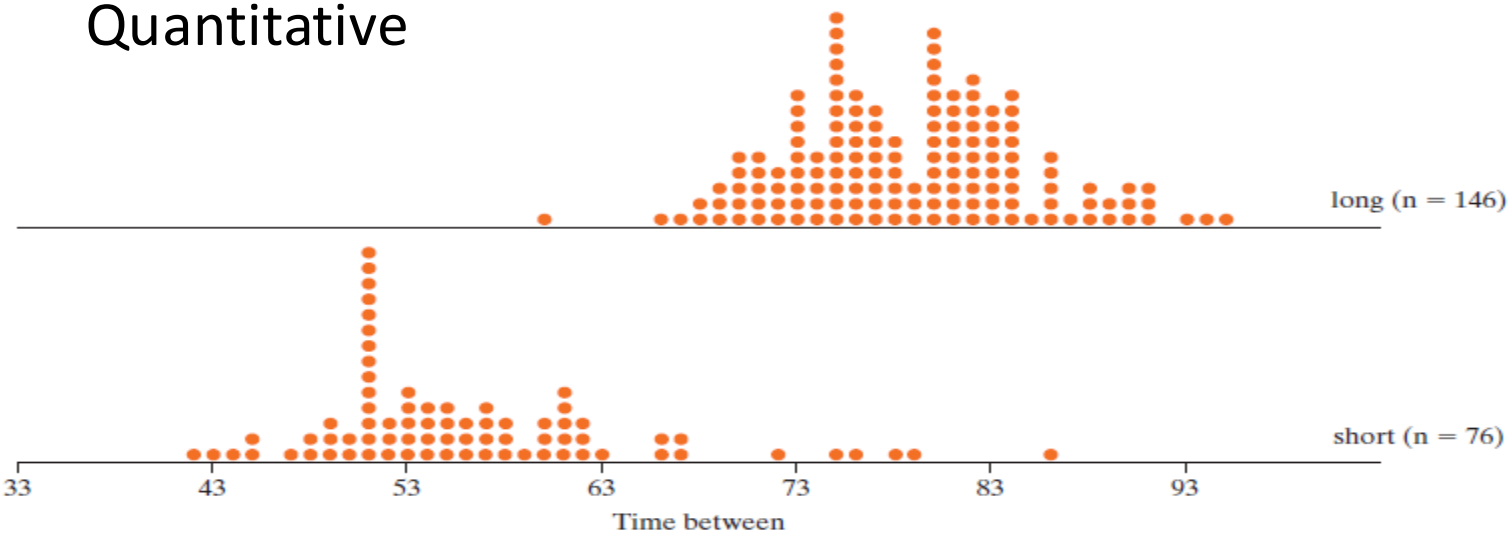
Dot Plot

Comparing Two Groups Graphically

Categorical



Quantitative



Notation Check

Statistics

- $\bar{x}$ Sample mean
- $\hat{p}$ Sample proportion.

Parameters

- μ Population mean
- π Population proportion or probability.

Statistics summarize a sample and parameters summarize a population

Quartiles

- Suppose 25% of the observations lie below a certain value x. Then x is called the ***lower quartile*** (or 25^{th} percentile).
- Similarly, if 25% of the observations are greater than x, then x is called the ***upper quartile*** (or 75^{th} percentile).
- The lower quartile can be calculated by finding the median, and then determining the median of the values below the overall median. Similarly the upper quartile is median$\{x_i : x_i >$ overall median$\}$.

IQR and Five-Number Summary

- The difference between the quartiles is called the ***inter-quartile range*** (IQR), another measure of variability along with standard deviation.
- The ***five-number summary*** for the distribution of a quantitative variable consists of the minimum, lower quartile, median, upper quartile, and maximum.
- Technically the IQR is not the interval (25th percentile, 75^{th} percentile), but the difference 75^{th} percentile – 25^{th} .
- Different software use different conventions, but we will use the convention that, if there is a range of possible quantiles, you take the middle of that range.
- For example, suppose data are 1, 3, 7, 7, 8, 9, 12, 14.
- M = 7.5, 25^{th} percentile = 5, 75^{th} percentile = 10.5. IQR = 5.5.

IQR and Five-Number Summary

- For medians and quartiles, we will use the convention, if there is a range of possibilities, take the middle of the range.
- In R, this is type = 2. type = 1 means take the minimum.
- x = c(1, 3, 7, 7, 8, 9, 12, 14)
- quantile(x,.25, type=2) ## 5
- IQR(x,type=2) ## 5.5
- IQR(x,type=1) ## 6. Can you see why?

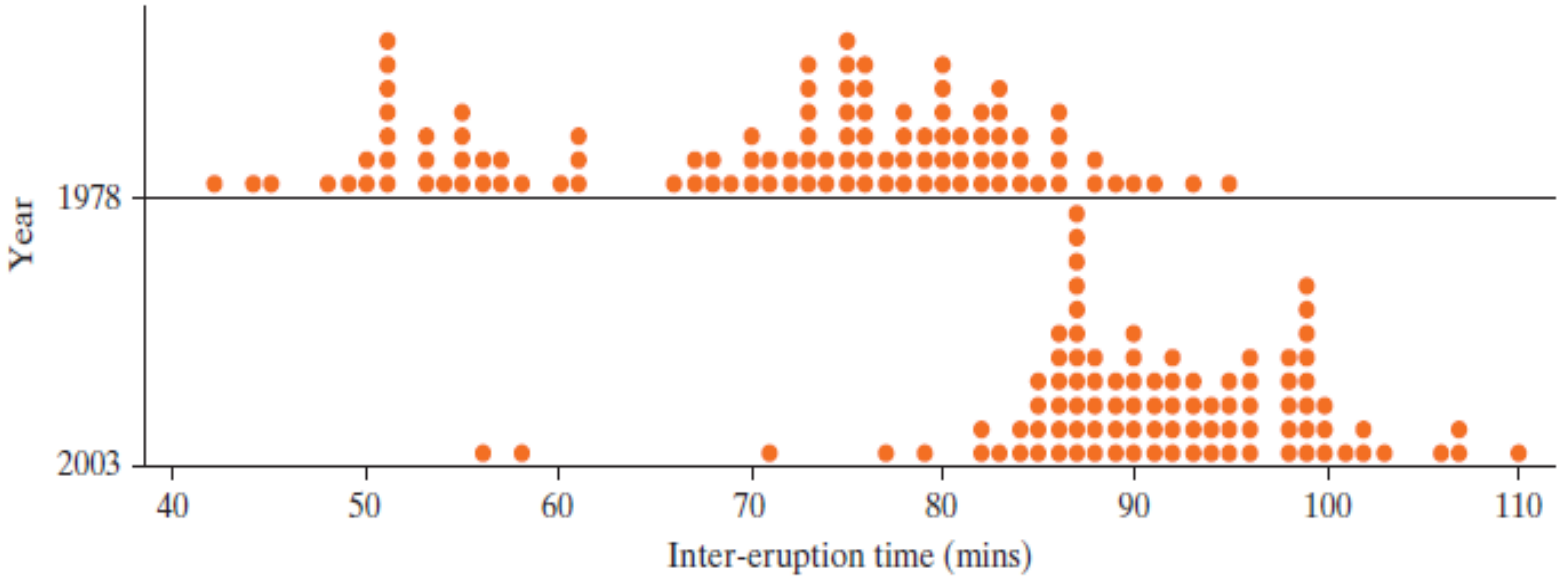
- For example, suppose data are 1, 3, 7, 7, 8, 9, 12, 14.
- M = 7.5, 25^{th} percentile = 5, 75^{th} percentile = 10.5. IQR = 5.5.

Geyser Eruptions

Example 6.1

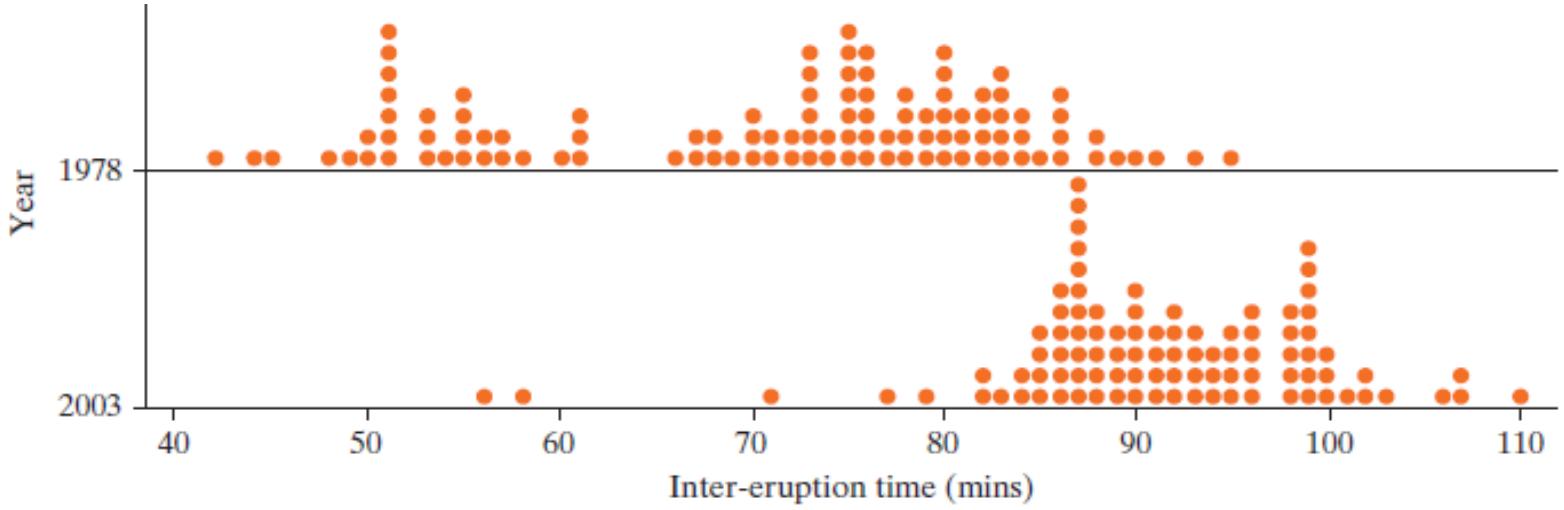
Old Faithful Inter-Eruption Times

- How do the five-number summary and IQR differ for inter-eruption times between 1978 and 2003?



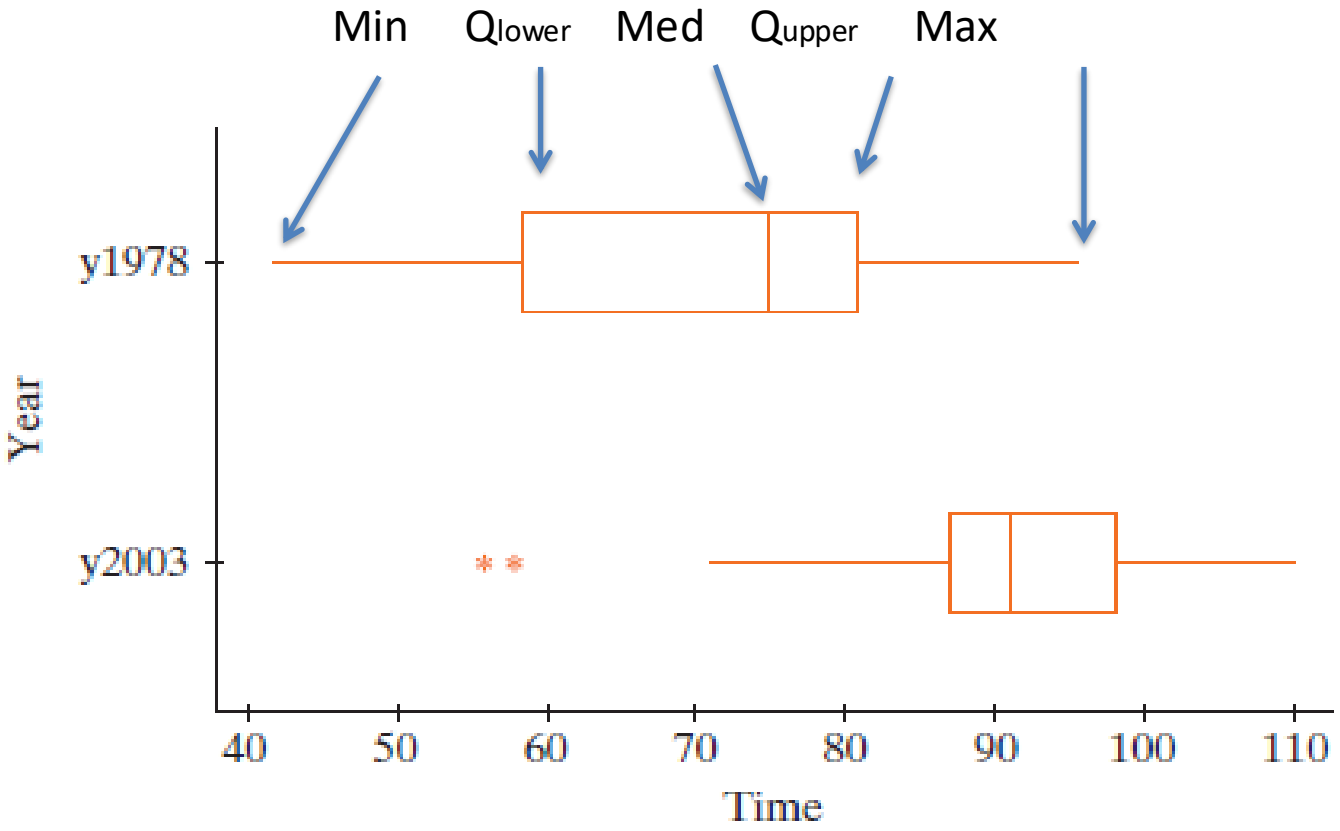
Old Faithful Inter-Eruption Times

	Minimum	Lower quartile	Median	Upper quartile	Maximum
1978 times	42	58	75	81	95
2003 times	56	87	91	98	110



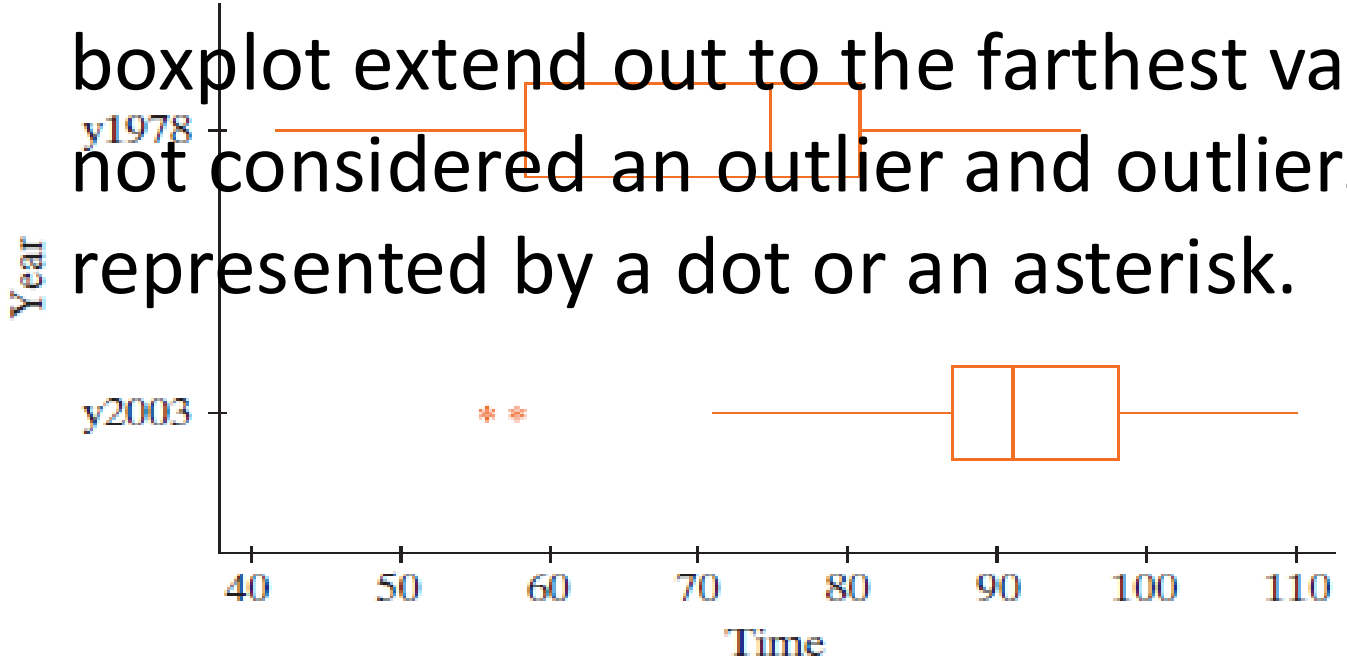
- 1978 IQR = 81 – 58 = 23
- 2003 IQR = 98 – 87 = 11

Boxplots


Min
Qlower
Med
Qupper
Max
y1978
y2003
Year
40
50
60
70
80
90
100
110
Time

Boxplots (Outliers)

- A data value that is more than 1.5 × IQR above the upper quartile or below the lower quartile is considered an outlier.
- When these occur, the whiskers on a boxplot extend out to the farthest value not considered an outlier and outliers are represented by a dot or an asterisk.

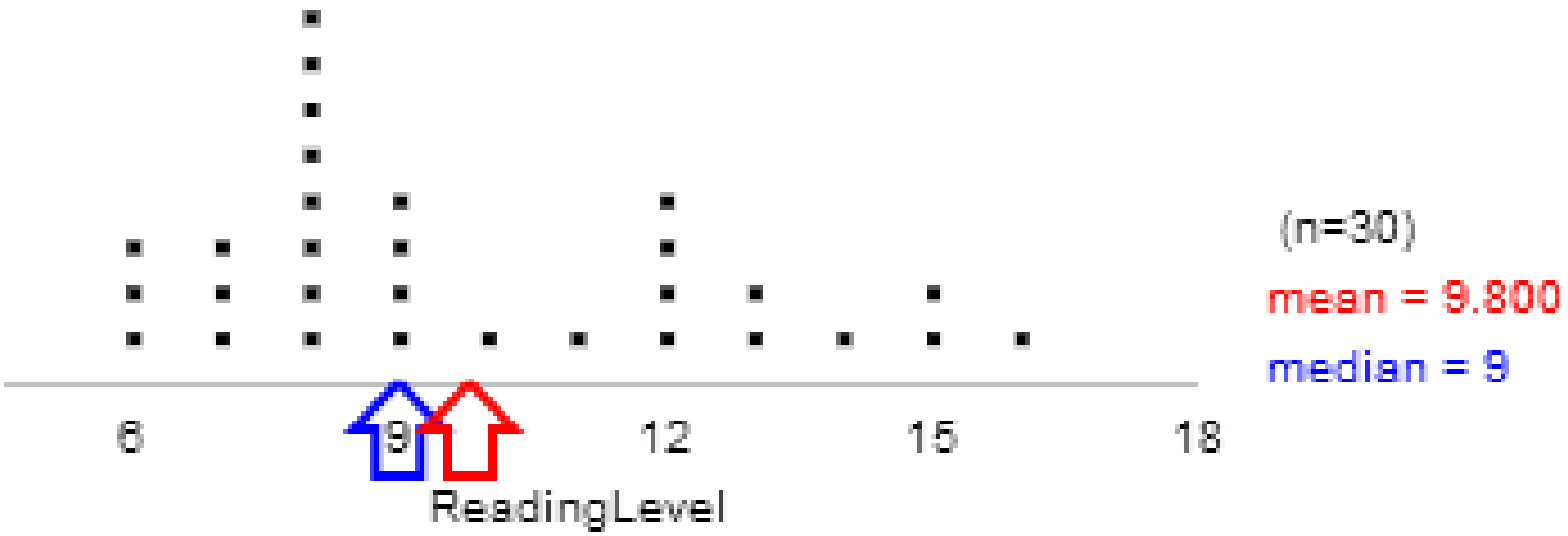


Cancer Pamphlet Reading Levels

- Short et al. (1995) compared reading levels of cancer patients and readability levels of cancer pamphlets. What is the:
  - Median reading level?
  - Mean reading level?
- Are the data skewed one way or the other?

Pamphlets' readability levels	6	7	8	9	10	11	12	13	14	15	16	Total
Count (number of pamphlets)	3	3	8	4	1	1	4	2	1	2	1	30

- Skewed a bit to the right
- Mean to the right of median



t-test, t CIs, and breastfeeding and intelligence example.

Example 6.3

Breastfeeding and Intelligence

- A 1999 study in *Pediatrics* examined if children who were breastfed during infancy differed from bottle-fed.
- 323 children recruited at birth in 1980-81 from four Western Michiga hospitals.
- Researchers deemed the participants representative of the communit social class, maternal education, age, marital status, and sex of infant
- Children were followed-up at age 4 and assessed using the General Cognitive Index (GCI)
  - A measure of the child's intellectual functioning
- Researchers surveyed parents and recorded if the child had been breastfed during infancy.

Breastfeeding and Intelligence

- Explanatory and response variables.
  - **Explanatory variable:** Whether the baby was breastfed. (Categorical)
  - **Response variable:** Baby's GCI at age 4. (Quantitative)

- Is this an experiment or an observational study?
- Can cause-and-effect conclusions be drawn in this study?

Breastfeeding and Intelligence

- **Null hypothesis:** There is no relationship between breastfeeding during infancy and GCI at age 4.
- **Alternative hypothesis:** There is a relationship between breastfeeding during infancy and GCI at age 4.

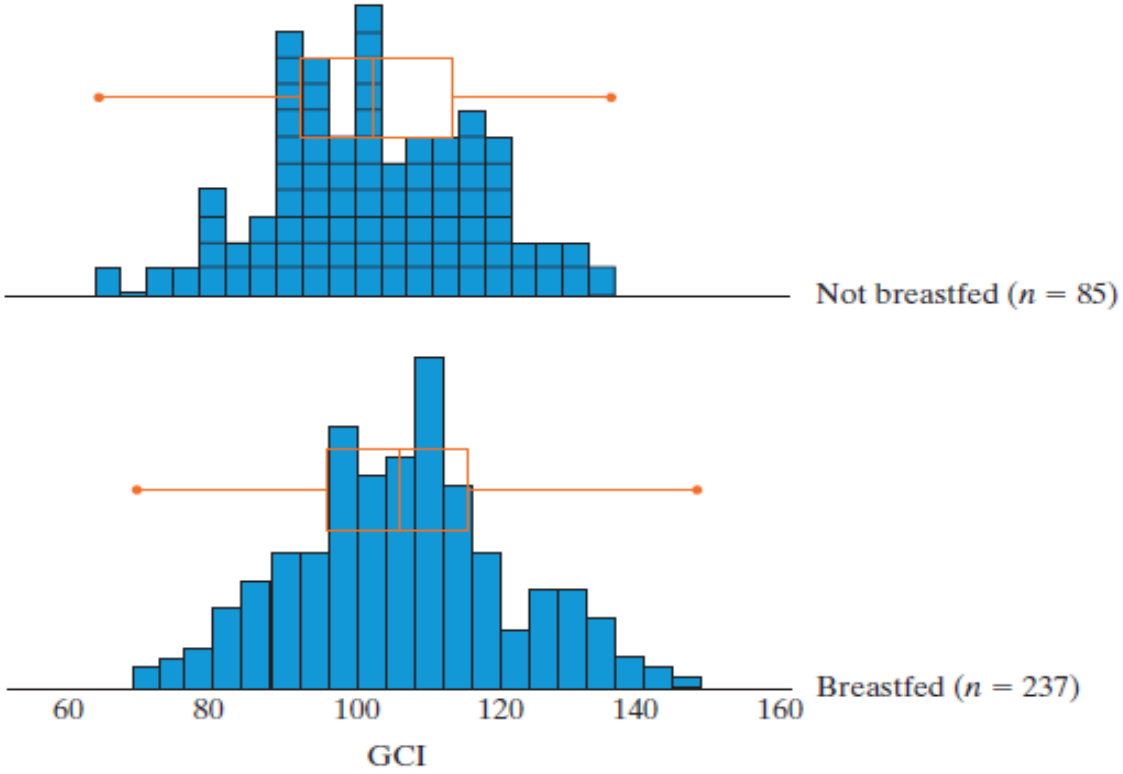
Breastfeeding and Intelligence

- $\mu_{breastfed}$ = Average GCI at age 4 for breastfed children
- μ_{not} = Average GCI at age 4 for children not breastfed

- **H_0:** $\mu_{breastfed} = \mu_{not}$
- **H_a:** $\mu_{breastfed} \neq \mu_{not}$

Breastfeeding and Intelligence

Group	Sample size, n	Sample mean	Sample SD
Breastfed	237	105.3	14.5
Not BF	85	100.9	14.0



Breastfeeding and Intelligence

The difference in means was 4.4.

- If breastfeeding is not related to GCI at age 4:
  - Is it **possible** a difference this large could happen by chance alone? Yes
  - Is it **plausible (believable, fairly likely)** a difference this large could happen by chance alone?
    - We can investigate this with simulations.
    - Alternatively, we can use a formula, or what your book calls a theory-based method.

T-statistic

- To use theory-based methods when comparing multiple means, the t-statistic is often used. Here the sample sizes are large, but if they were small and the populations were normal, the t-test would be more appropriate than the z-test.
- the t-statistic is again simply the number of standard errors our statistic is above or below the mean under the null hypothesis.
- $t = \frac{statistic - hypothesized\ value\ under\ Ho}{SE} = \frac{\bar{x}_1 - \bar{x}_2 - 0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$
- Here, t = $\frac{(105.3 - 100.9) - 0}{\sqrt{(\frac{14.5^2}{237} + \frac{14.0^2}{85})}} = 2.46.$
- p-value ~ 1.4 or 1.5%. [2 * (1-pnorm(2.46))], or use pt.