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

1. Comparing two proportions using formulas, smoking and gender, continued.
2. Five number summary and IQR.
3. t-test for comparing two means.
4. t versus normal, and when to use what formula.
5. Causation and prediction.
6. Review list.
7. Example problems.

Read ch5 and 6. The midterm will be on ch 1-6.
Hw3 is due Thu Aug31, 10am, to statgrader or stagrader2.
Midterm is Thu Aug31, 10am.
For the midterm, you will have from 10am to 11:20am.
I will post the midterm on the course website,
http://www.stat.ucla.edu/~frederic/13/sum23 .
First see the file MidtermInstructions.txt which will be posted there Thu at 9am.
You also need to zoom in to the usual zoom while taking the exam.
By 11:20am you must email me your answers, to frederic@stat.UCLA.edu .
After the exam there will be no lecture.
Your email should just contain your answers, like
ADDBC CDAAB BBCCA .
If you forsee possible internet problems, submit you answers a few min early!!!

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

- The margin of error for the difference in proportions is

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

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

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

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

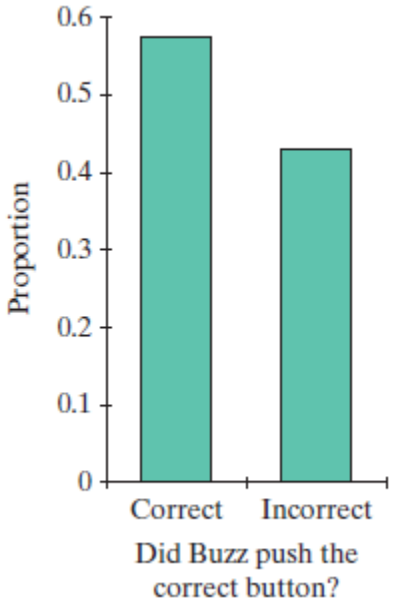
Section 6.1

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…

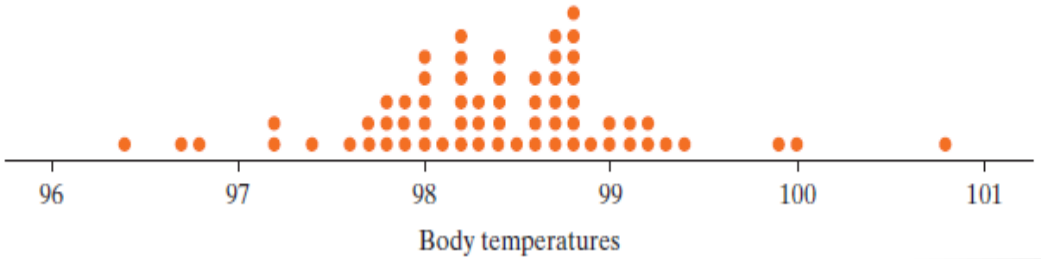
Graphs for a Single Variable

Categorical



Bar Graph

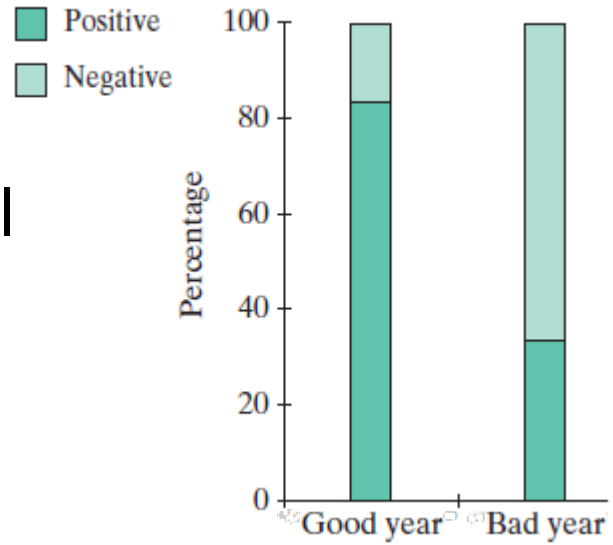
Quantitative



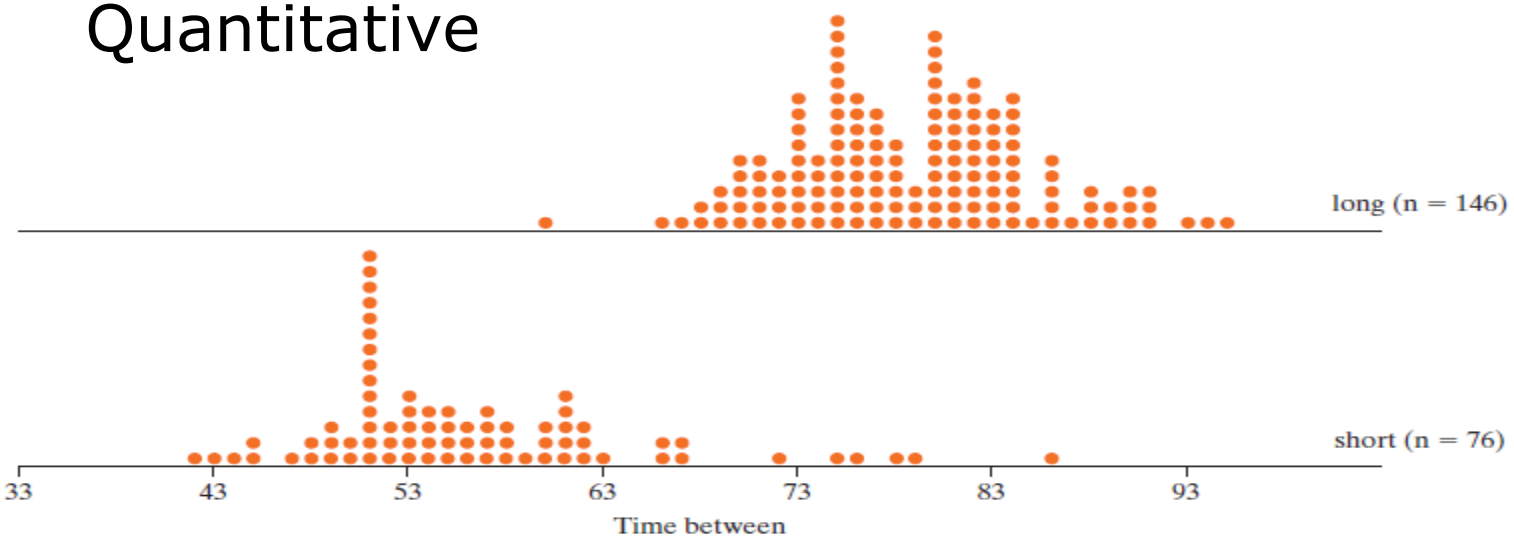
Dot Plot

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.

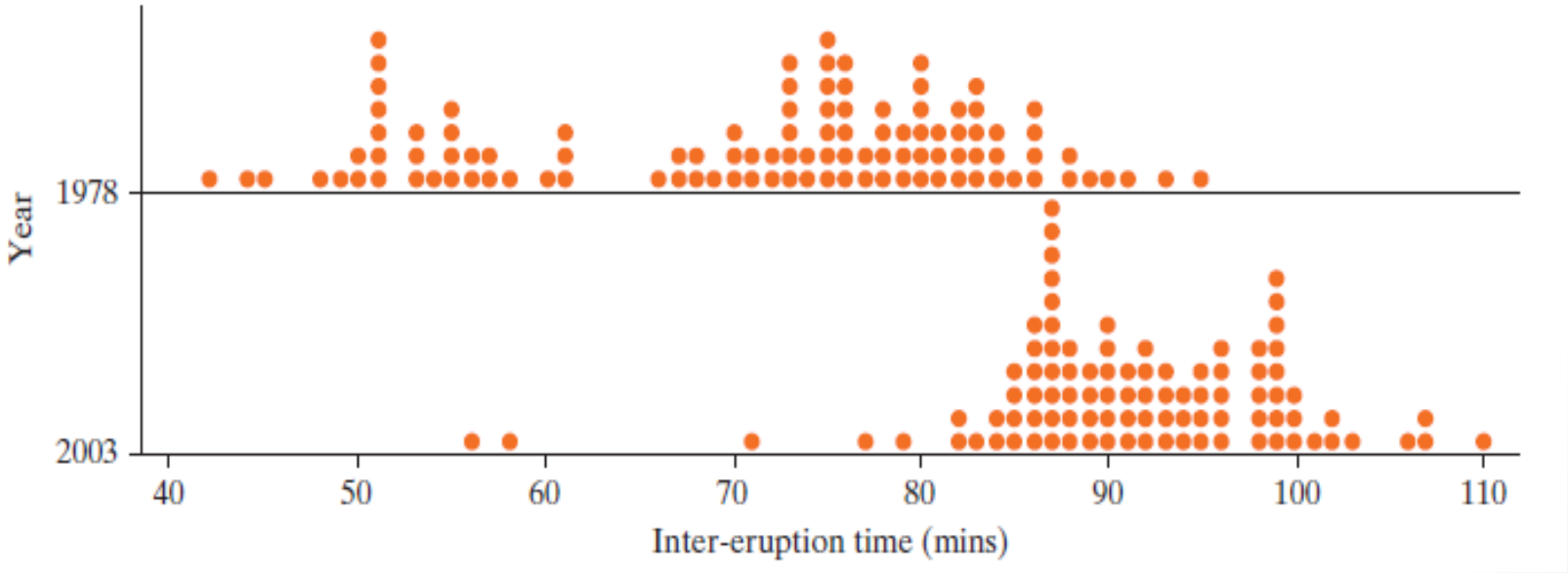
IQR and Five-Number Summary

- 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) ## 5.5
- 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, 25th percentile = 5, 75th percentile = 10.5. IQR = 5.5.

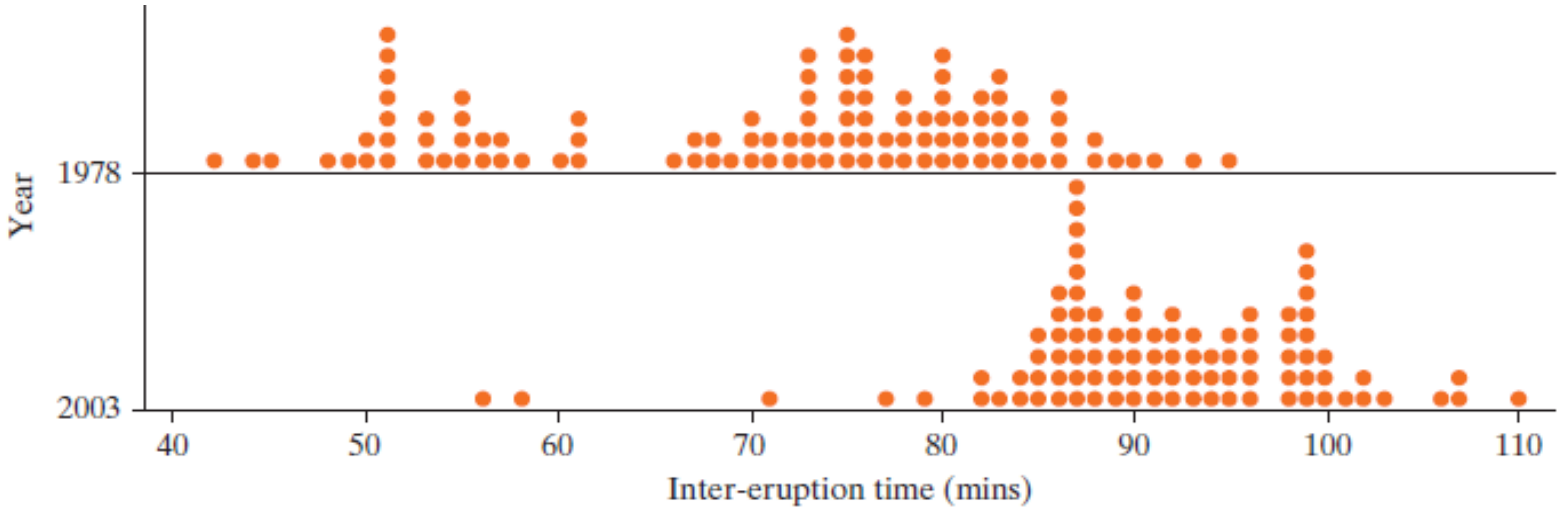
Old Faithful Inter-Eruption Times

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



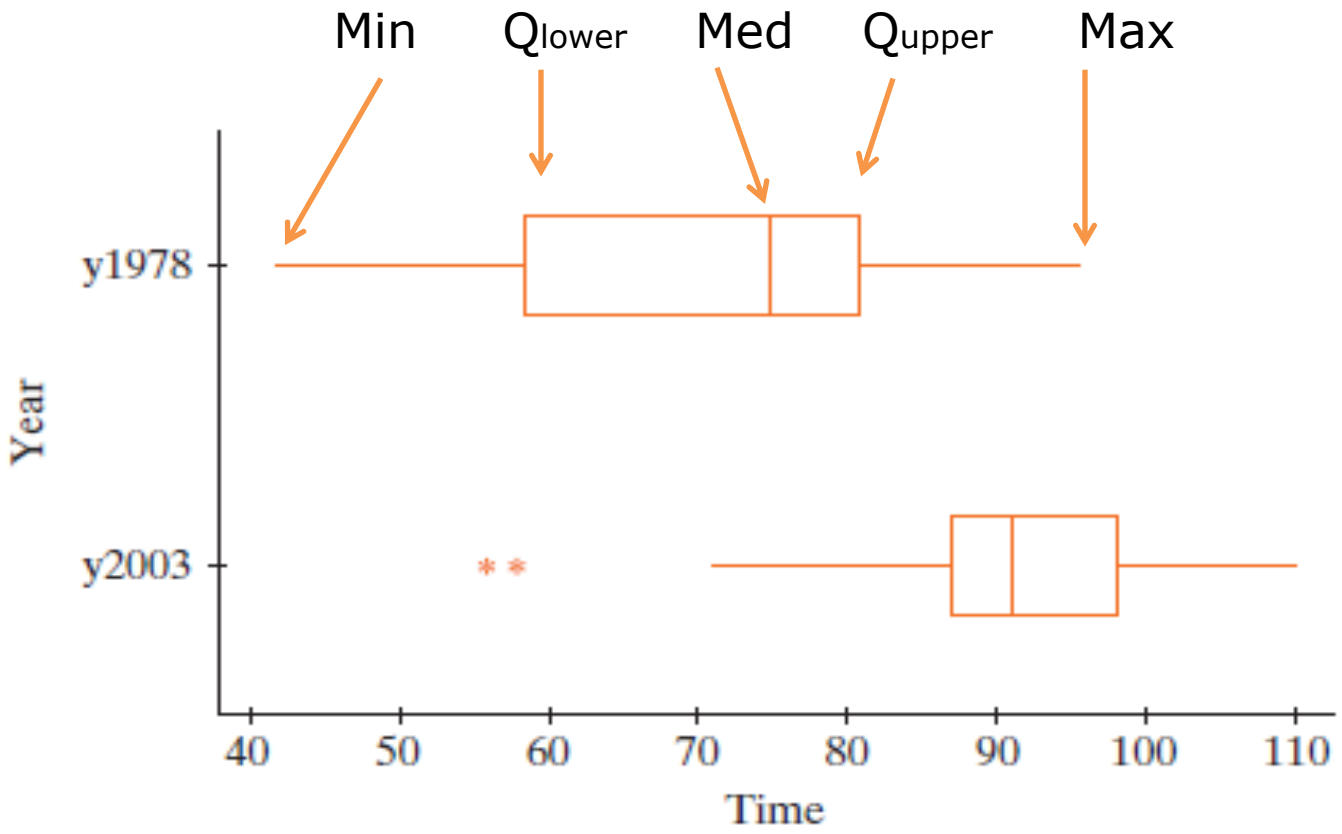
Old Faithful Inter-Eruption Times

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



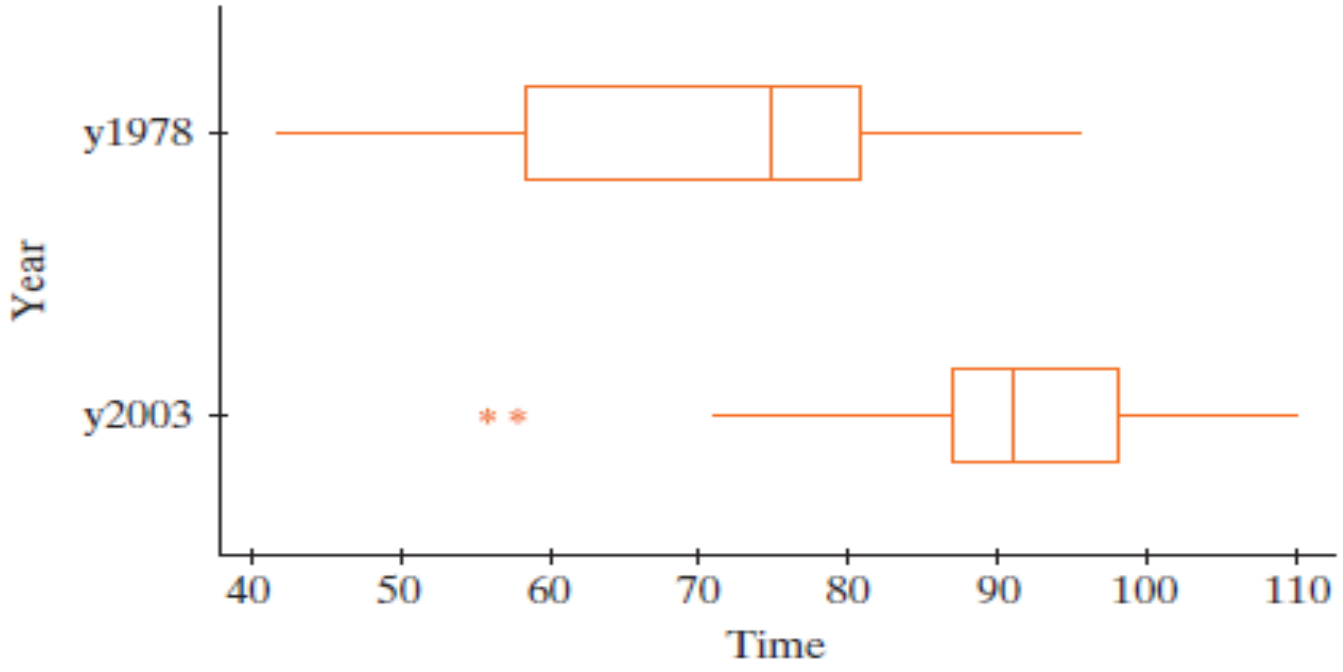
- 1978 IQR = 81 – 58 = 23
- 2003 IQR = 98 – 87 = 11

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.

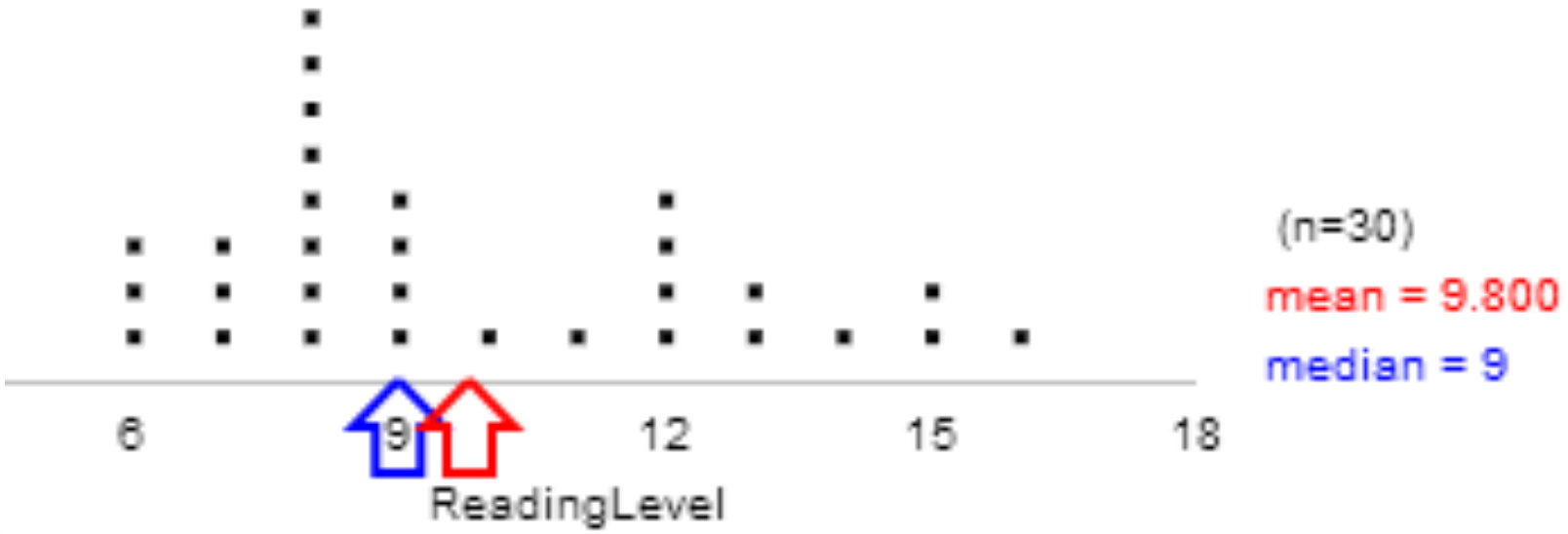


Pamphlet Reading Levels

- 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 Michigan hospitals.
- Researchers deemed the participants representative of the community in 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.

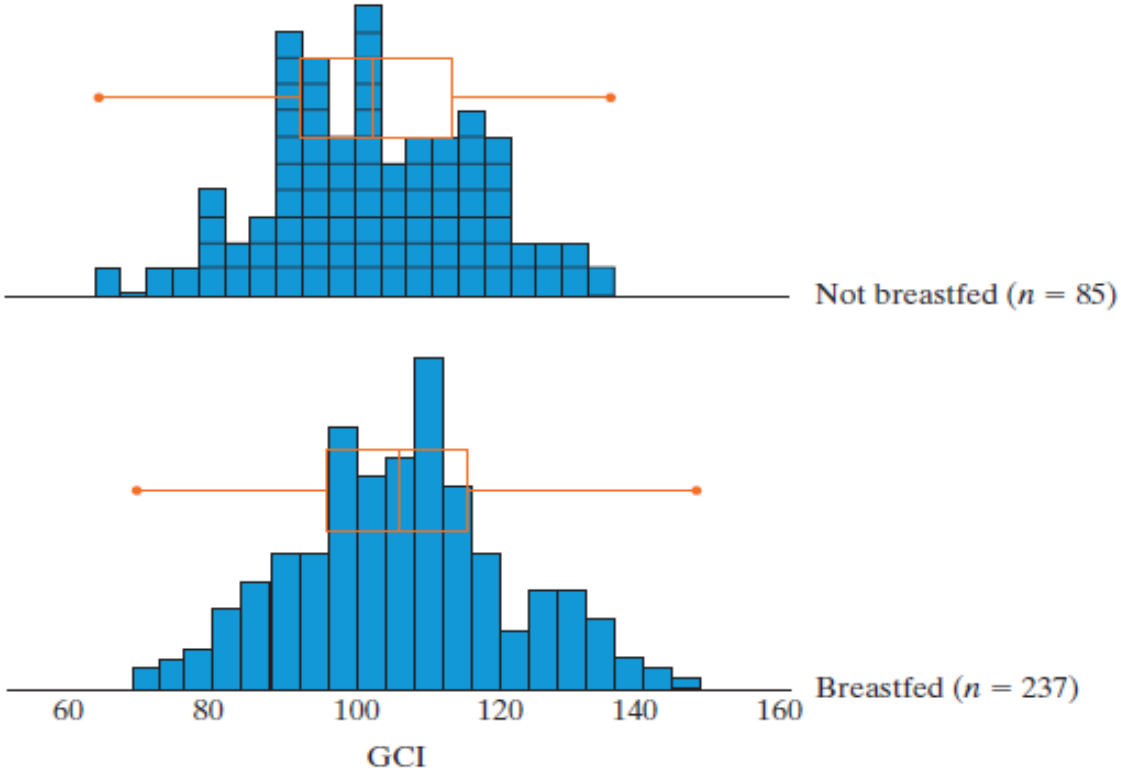
Breastfeeding and Intelligence

- $\mu_{breastfed}$ = 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.

Breastfeeding and Intelligence

Meaning of the p-value:

- If breastfeeding were not related to GCI at age 4, then the probability of observing a difference of 4.4 or more or -4.4 or less just by chance is about 1.4%.

- A 95% CI can also be obtained using the t-distribution. The SE is $\sqrt{(\frac{14.5^2}{237} + \frac{14.0^2}{85})}$ = 1.79. So the margin of error is multiplier x SE.

Breastfeeding and Intelligence

- The SE is $\sqrt{(\frac{14.5^2}{237} + \frac{14.0^2}{85})}$ = 1.79. The margin of error is multiplier x SE.
- The multiplier should technically be obtained using the t distribution, but for large sample sizes you get almost the same multiplier with t and normal. Use 1.96 for a 95% CI to get 4.40 +/- 1.96 x 1.79 = 4.40 +/- 3.51 = (0.89, 7.91).
- The book uses 2 instead of 1.96, and the applet uses 1.9756 from the t-distribution. Just use 1.96 for 95% CIs for this class.

Breastfeeding and Intelligence

- We have strong evidence against the null hypothesis and can conclude the association between breastfeeding and intelligence here is statistically significant.
- Breastfed babies have statistically significantly higher average GCI scores at age 4.
- We can see this in both the small p-value (0.015) and the confidence interval that says the mean GCI for breastfed babies is 0.89 to 7.91 points higher than that for non-breastfed babies.

Breastfeeding and Intelligence

- Can you conclude that breastfeeding improves average GCI at age 4?
  - No. The study was not a randomized experiment.
  - We cannot conclude a cause-and-effect relationship.
- There might be alternative explanations for the significant difference in average GCI values.
- What might some confounding factors be?

Breastfeeding and Intelligence

- Can you conclude that breastfeeding improves average GCI at age 4?
  - No. The study was not a randomized experiment.
  - We cannot conclude a cause-and-effect relationship.
- There might be alternative explanations for the significant difference in average GCI values.
  - Maybe better educated mothers are more likely to breastfeed their children
  - Maybe mothers that breastfeed spend more time with their children and interact with them more.
  - Some mothers who do not breastfeed are less healthy or their babies have weaker appetites and this might slow down development in general.

t versus normal, and when to use what formula.

Why do we sometimes use the t distribution and sometimes the normal distribution in testing and confidence intervals?

The central limit theorem states that, for any iid random variables X_1, ..., X_n with mean μ and SD σ, $(\bar{x} - \mu) \div (\sigma/\sqrt{n})$ -> standard normal, as n -> ∞.

iid means independent and identically distributed, like draws from the same large population.

standard means mean 0 and SD 1.

t versus normal and assumptions.

CLT: $(\bar{x} - \mu) \div (\sigma/\sqrt{n})$ -> standard normal.

If Z is std. normal, then $P(|Z| < 1.96) = 95\%$.

So, if n is large, then

$P(|(\bar{x} - \mu) \div (\sigma/\sqrt{n})| < 1.96) \sim 95\%$.

Mult. by $(\sigma/\sqrt{n})$ and get

$P(|\bar{x} - \mu| < 1.96\ \sigma/\sqrt{n}) \sim 95\%$.

$P(\mu - \bar{x}$ is in the range 0 +/- $1.96\ \sigma/\sqrt{n}) \sim 95\%$.

$P(\mu$ is in the range $\bar{x}$ +/- $1.96\ \sigma/\sqrt{n}) \sim 95\%$.

This all assumes n is large. What if n is small?

t versus normal and assumptions.

CLT: $(\bar{x} - \mu) \div (\sigma/\sqrt{n})$ -> standard normal.

What about if n is small?

A property of the normal distribution is that the sum of independent normals is also normal, and from this it follows that if $X_1, \ldots, X_n$ are iid and normal, then $(\bar{x} - \mu) \div (\sigma/\sqrt{n})$ is standard normal.

So again P(μ is in the range $\bar{x}$ +/- 1.96 $\sigma/\sqrt{n}$) = 95%.

This assumes you know σ. What if σ is unknown?

t versus normal and assumptions.

Suppose $X_1, \ldots, X_n$ are iid with mean μ and SD σ.

CLT: $(\bar{x} - \mu) \div (\sigma/\sqrt{n}) \sim$ std. normal.

If $X_1, \ldots, X_n$ are normal, then $(\bar{x} - \mu) \div (\sigma/\sqrt{n})$ is std. normal.

σ is the SD of the population from which $X_1, \ldots, X_n$ are drawn. s is the SD of the sample, $X_1, \ldots, X_n$.

Gosset (1908) showed that replacing σ with s,

if $X_1, \ldots, X_n$ are normal, then $(\bar{x} - \mu) \div (s/\sqrt{n})$ is t distributed.

So we need the multiplier from the t distribution.

t versus normal and assumptions.

To sum up,

if the observations are iid and n is large, then

P(μ is in the range $\bar{x}$ +/- 1.96 σ/√n) ~ 95%.

If the observations are iid and normal, then

P(μ is in the range $\bar{x}$ +/- 1.96 σ/√n) ~ 95%.

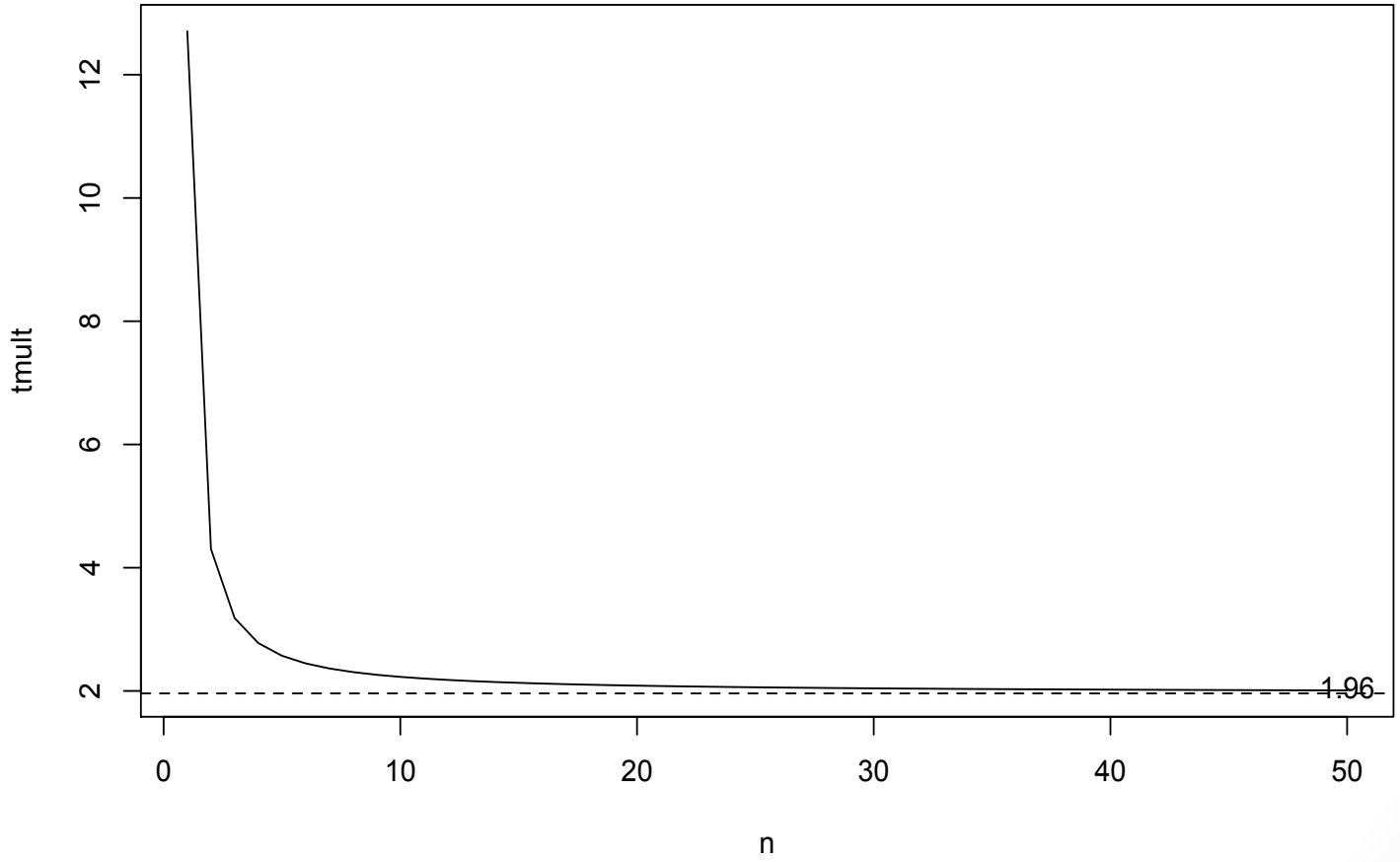
If the obs. are iid and normal and σ is unknown, then

P(μ is in the range $\bar{x}$ +/- t_{mult} s/√n) ~ 95%.

where t_{mult} is the multiplier from the t distribution.

This multiplier depends on n.

t versus normal and assumptions.



When to use which formula.

a. 1 sample numerical data, iid observations, want a 95% CI for μ.

- If n is large and σ is known, use $\bar{x}$ +/- 1.96 $\sigma/\sqrt{n}$.
- If n is small, draws are normal, and σ is known, use $\bar{x}$ +/- 1.96 $\sigma/\sqrt{n}$.
- If n is small, draws are normal, and σ is unknown, use $\bar{x}$ +/- t_{mult} $s/\sqrt{n}$.
- If n is large and σ is unknown, t_{mult} ~ 1.96, so we can use $\bar{x}$ +/- 1.96 $s/\sqrt{n}$.

n ≥ 30 is often considered large enough to use 1.96.

In practice, we typically do not know the draws are normal, but if the distribution looks roughly symmetrical without enormous outliers, the t formula may be reasonable.

b. 1 sample binary data, iid observations, want a 95% CI for π.

View the data as 0 or 1, so sample percentage p = $\bar{x}$, and

s = $\sqrt{[p(1-p)]}$, $\sigma = \sqrt{[\pi(1-\pi)]}$.

When to use which formula.

a. 1 sample numerical data, iid observations, want a 95% CI for μ.

- If n is large and σ is known, use $\bar{x}$ +/- 1.96 $\sigma/\sqrt{n}$.
- If n is small, draws are normal, and σ is known, use $\bar{x}$ +/- 1.96 $\sigma/\sqrt{n}$.
- If n is small, draws ~ normal, and σ is unknown, use $\bar{x}$ +/- t_{mult} $s/\sqrt{n}$.
- If n is large and σ is unknown, t_{mult} ~ 1.96, so we can use $\bar{x}$ +/- 1.96 s/√n.

b. 1 sample binary data, iid observations, want a 95% CI for π.

View the data as 0 or 1, so sample percentage p = $\bar{x}$, and

s = √[p(1-p)], $\sigma = \sqrt{[\pi(1-\pi)]}$.

If n is large and π is unknown, use $\bar{x}$ +/- 1.96 s/√n.

Here large n means ≥ 10 of each type in the sample.

When to use which formula.

What if n is small and the draws are not normal, and you want a theory-based test or CI?

How should you find the t multiplier for a CI or a p-value using the t-statistic, when n is small?

These are questions outside the scope of this course, but some techniques have been developed, such as the bootstrap, which are sometimes useful in these situations.

When to use which formula.

c. Numerical data from 2 samples, iid observations, want a 95% CI for $\mu_1 - \mu_2$.

If n is large and σ is unknown, use $\bar{x}_1 - \bar{x}_2 +/- 1.96 \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$.

As with one sample, if σ_1 is known, replace s_1 with σ_1, and the same for σ_2. And as with one sample, if σ_1 and σ_2 are unknown, the sample sizes are small, and the distributions are roughly normal, then use t_{mult} instead of 1.96. If the sample sizes are small, the distributions are normal, and σ_1 and σ_2 are known, then use 1.96.

d. Binary data from 2 samples, iid observations, want a 95% CI for $\pi_1 - \pi_2$.

same as in c above, with $p_1 = \bar{x}_1$, $s_1 = \sqrt{[p_1 (1-p_1)]}$, $\sigma_1 = \sqrt{[\pi_1 (1-\pi_1)]}$.

Large for binary data means sample has ≥ 10 of each type.

Causation and prediction.

Note that for prediction, you sometimes do not care about confounding factors.

* Forecasting wildfire activity using temperature.

Warmer weather may directly cause wildfires via increased ease of ignition, or due to confounding with people choosing to go camping in warmer weather. It does not really matter for the purpose of merely *predicting* how many wildfires will occur in the coming month.

* The same goes for predicting lifespan, or liver disease rates, etc., using smoking as a predictor variable.

Review list.

1. Meaning of SD.
2. Parameters and statistics.
3. Z statistic for proportions.
4. Simulation and meaning of pvalues.
5. SE for proportions.
6. What influences pvalues.
7. CLT and validity conditions for tests.
8. 1-sided and 2-sided tests.
9. Reject the null vs. accept the alternative.
10. Sampling and bias.
11. Significance level.
12. Type I, type II errors, and power.
13. CIs for a proportion.
14. CIs for a mean.
15. Margin of error.
16. Practical significance.
17. Confounding.
18. Observational studies and experiments.
19. Random sampling and random assignment.
20. Two proportion CIs and testing.
21. IQR and 5 number summaries.
22. Testing and CIs for 2 means.
23. Placebo effect, adherer bias, and nonresponse bias.
24. Prediction and causation.

Example problems.

NCIS was the top-rated tv show in 2014. It was 3rd in 2016 and is now 5th in 2017.

A study finds that in a certain city, people who watch NCIS are much more likely to die than people who do not watch NCIS. Can we conclude that NCIS is a dangerous tv show to watch?

Example problems.

NCIS was the top-rated tv show in 2022.

A study finds that in a certain city, people who watch NCIS are much more likely to die than people who do not watch NCIS. Can we conclude that NCIS is a dangerous tv show to watch?

No. Age is a confounding factor. The median age of a viewer is 61 years old.

- 1. Suppose the population of American adults has a mean systolic blood pressure of 120 mm Hg and an SD of 20 mm Hg. You take a simple random sample of 100 American adults. Which of the following is true?
- A typical adult's blood pressure would differ from 120 by about **20** mm Hg, and a typical sample of size 100 would have a sample mean that differs from 120 by about **2** mm Hg.
- A typical adult's blood pressure would differ from 120 by about **20** mm Hg, and a typical sample of size 100 would have a sample mean that differs from 120 by about **20** mm Hg.
- A typical adult's blood pressure would differ from 120 by about **2** mm Hg, and a typical sample of size 100 would have a sample mean that differs from 120 by about **0.2** mm Hg.
- A typical adult's blood pressure would differ from 120 by about **20** mm Hg, and a typical sample of size 100 would have a sample mean that differs from 120 by about **0.2** mm Hg.

- 1. Suppose the population of American adults has a mean systolic blood pressure of 120 mm Hg and an SD of 20 mm Hg. You take a simple random sample of 100 American adults. Which of the following is true?
- **A typical adult's blood pressure would differ from 120 by about 20 mm Hg, and a typical sample of size 100 would have a sample mean that differs from 120 by about 2 mm Hg.**
- A typical adult's blood pressure would differ from 120 by about **20** mm Hg, and a typical sample of size 100 would have a sample mean that differs from 120 by about **20** mm Hg.
- A typical adult's blood pressure would differ from 120 by about **2** mm Hg, and a typical sample of size 100 would have a sample mean that differs from 120 by about **0.2** mm Hg.
- A typical adult's blood pressure would differ from 120 by about **20** mm Hg, and a typical sample of size 100 would have a sample mean that differs from 120 by about **0.2** mm Hg.

EXAMPLE PROBLEMS.

- In the study on echinacea by O'Neil et al. (2008), with 58 volunteers, which of the following is a valid conclusion based on the data?
- a. This was an observational study, so confounding factors such as the overall health and wealth of the volunteers are a likely explanation for the results.
- b. In this experiment, the echinacea group got sick less than the placebo group, but the difference was not statistically significant, in part because the sample size was so small.
- c. This study showed that echinacea works to prevent colds, but its effect is very minimal.
- d. The explanatory variable is a confounding factor t-test with 95% central limit theorem.
- e. None of the above.

EXAMPLE PROBLEMS.

- In the study on echinacea by O'Neil et al. (2008), with 58 volunteers, which of the following is a valid conclusion based on the data?
- a. This was an observational study, so confounding factors such as the overall health and wealth of the volunteers are a likely explanation for the results.
- **b. In this experiment, the echinacea group got sick less than the placebo group, but the difference was not statistically significant, in part because the sample size was so small.**
- c. This study showed that echinacea works to prevent colds, but its effect is very minimal.
- d. The explanatory variable is a confounding factor t-test with 95% central limit theorem.
- e. None of the above.

Example problems.

Suppose you sample 100 UCLA students and 80 2nd graders and record their blood glucose levels, to see if going to UCLA is associated with higher levels. The mean at UCLA is 5.5 mmol/L, with an SD of 1.5, and the mean in 2nd grade is 7.5 mmol/L, with an SD of 2.2.

a. Find a 95%-CI for how much less an average UCLA student's blood glucose level is than an average 2nd grader.

Example problems.

Suppose you sample 100 UCLA students and 80 2nd graders and record their blood glucose levels. The mean at UCLA is 5.5 mmol/L, with an SD of 1.5, and the mean in 2nd grade is 7.5 mmol/L, with an SD of 2.2.

a. Find a 95%-CI for how much less an average UCLA student's blood glucose level is than an average 2nd grader.

$2.0 +/- 1.96 \sqrt{(1.5^2/100 + 2.2^2/80)} = 2.0 +/- 0.564.$

Example problems.

Suppose you sample 100 UCLA students and 80 2nd graders and record their blood glucose levels. The mean at UCLA is 5.5 mmol/L, with an SD of 1.5, and the mean in 2nd grade is 7.5 mmol/L, with an SD of 2.2.

b. Is the difference observed between the mean blood glucose at UCLA and in 2nd grade statistically significant?

Example problems.

Suppose you sample 100 UCLA students and 80 2nd graders and record their blood glucose levels. The mean at UCLA is 5.5 mmol/L, with an SD of 1.5, and the mean in 2nd grade is 7.5 mmol/L, with an SD of 2.2.

b. Is the difference observed between the mean blood glucose at UCLA and in 2nd grade statistically significant?

Yes. The 95%-CI does not come close to containing 0.

Example problems.

Suppose you sample 100 UCLA students and 80 2nd graders and record their blood glucose levels. The mean at UCLA is 5.5 mmol/L, with an SD of 1.5, and the mean in 2nd grade is 7.5 mmol/L, with an SD of 2.2.

c. Is this an observational study or an experiment?

Example problems.

Suppose you sample 100 UCLA students and 80 2nd graders and record their blood glucose levels. The mean at UCLA is 5.5 mmol/L, with an SD of 1.5, and the mean in 2nd grade is 7.5 mmol/L, with an SD of 2.2.

c. Is this an observational study or an experiment?

Observational study.

Example problems.

Suppose you sample 100 UCLA students and 80 2nd graders and record their blood glucose levels. The mean at UCLA is 5.5 mmol/L, with an SD of 1.5, and the mean in 2nd grade is 7.5 mmol/L, with an SD of 2.2.

d. Does going to UCLA cause your blood glucose level to drop?

Example problems.

Suppose you sample 100 UCLA students and 80 2nd graders and record their blood glucose levels. The mean at UCLA is 5.5 mmol/L, with an SD of 1.5, and the mean in 2nd grade is 7.5 mmol/L, with an SD of 2.2.

d. Does going to UCLA cause your blood glucose level to drop?

No. Age is a confounding factor. Young kids eat more candy.

Example problems.

Suppose you sample 100 UCLA students and 80 2nd graders and record their blood glucose levels. The mean at UCLA is 5.5 mmol/L, with an SD of 1.5, and the mean in 2nd grade is 7.5 mmol/L, with an SD of 2.2.

e. The mean blood glucose level of all 43,301 UCLA students is a

parameter random variable t-test

Example problems.

Suppose you sample 100 UCLA students and 80 2nd graders and record their blood glucose levels. The mean at UCLA is 5.5 mmol/L, with an SD of 1.5, and the mean in 2nd grade is 7.5 mmol/L, with an SD of 2.2.

e. The mean blood glucose level of all 43,301 UCLA students is a

parameter

Example problems.

Suppose you sample 100 UCLA students and 80 2nd graders and record their blood glucose levels. The mean at UCLA is 5.5 mmol/L, with an SD of 1.5, and the mean in 2nd grade is 7.5 mmol/L, with an SD of 2.2.

f. If we took another sample of 100 UCLA students and 80 2nd graders, and used the difference in sample means to estimate the difference in population means, how much would it typically be off by?

Example problems.

Suppose you sample 100 UCLA students and 80 2nd graders and record their blood glucose levels. The mean at UCLA is 5.5 mmol/L, with an SD of 1.5, and the mean in 2nd grade is 7.5 mmol/L, with an SD of 2.2.

f. If we took another sample of 100 UCLA students and 80 2nd graders, and used the difference in sample means to estimate the difference in population means, how much would it typically be off by? $SE = \sqrt{(1.5^2/100 + 2.2^2/80)} = .288$ mmol/L

Example problems.

Suppose you sample 100 UCLA students and 80 2nd graders and record their blood glucose levels. The mean at UCLA is 5.5 mmol/L, with an SD of 1.5, and the mean in 2nd grade is 7.5 mmol/L, with an SD of 2.2.

g. How much does one UCLA student's blood glucose level typically differ from the mean of UCLA students?

Example problems.

Suppose you sample 100 UCLA students and 80 2nd graders and record their blood glucose levels. The mean at UCLA is 5.5 mmol/L, with an SD of 1.5, and the mean in 2nd grade is 7.5 mmol/L, with an SD of 2.2.

g. How much does one UCLA student's blood glucose level typically differ from the mean of UCLA students?

1.5 mmoL/L.

Example problems.

- Researchers take a simple random sample of Californians and a simple random sample of Texans to see who does more exercise. They find that the Californians spend 2.5 hours per week exercising on average and the Texans spend 2.0 hours per week exercising on average. The researchers do a 2-sided test on the difference between the two means and find a p-value of 2.3%. Which of the following would be true of 90% and 95% confidence intervals for the weekly mean exercising time for Californians minus the mean exercising time for Texans?
- a. Both the 90% CI and the 95% CI will contain zero.
- b. Neither the 90% CI nor the 95% CI will contain zero.
- c. The 95% CI will not contain zero, but the 90% CI might contain zero.
- d. The 95% CI will contain zero, but the 90% CI might not contain zero.

Example problems.

- Researchers take a simple random sample of Californians and a simple random sample of Texans to see who does more exercise. They find that the Californians spend 2.5 hours per week exercising on average and the Texans spend 2.0 hours per week exercising on average. The researchers do a 2-sided test on the difference between the two means and find a p-value of 2.3%. Which of the following would be true of 90% and 95% confidence intervals for the weekly mean exercising time for Californians minus the mean exercising time for Texans?
- a. Both the 90% CI and the 95% CI will contain zero.
- **b. Neither the 90% CI nor the 95% CI will contain zero.**
- c. The 95% CI will not contain zero, but the 90% CI might contain zero.
- d. The 95% CI will contain zero, but the 90% CI might not contain zero.

Example problems.

- The Physician's Health Study I studied aspirin's effect on reducing the risk of heart attacks. Which of the following was **not** a reason for randomly assigning people to treatment or control in this experiment?
- a. To ensure that the sample is more representative of the overall population.
- b. To ensure that the treatment and control groups are similar with respect to known potential confounders such as diet and exercise.
- c. To ensure that the treatment and control groups are similar with respect to unknown confounding factors.

Example problems.

- The Physician's Health Study I studied aspirin's effect on reducing the risk of heart attacks. Which of the following was **not** a reason for randomly assigning people to treatment or control in this experiment?
- **a. To ensure that the sample is more representative of the overall population.**
- b. To ensure that the treatment and control groups are similar with respect to known potential confounders such as diet and exercise.
- c. To ensure that the treatment and control groups are similar with respect to unknown confounding factors.