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

1. Five number summary, IQR, boxplots, and geysers example.
2. t-test, t CIs, and breastfeeding and intelligence example.
3. Causation and prediction.
4. When to use which formula.

http://www.stat.ucla.edu/~frederic/13/W24 .
HW2 due today, Feb12, 1159pm. 2.3.15, 3.3.18, and 4.1.23.
No class or OH Mon Feb19, President's Day.
Midterm exam from a previous year is on the course website, in the file old13midterm.pdf .
Read ch5-6. The midterm will be on ch 1-6.
Midterm is Mon Feb26 in class. Bring a pencil or pen, and a calculator.
On the exam, you cannot use computers or ipads or phones or anything that can surf the web or do email.

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, 75th percentile), but the difference 75th percentile – 25th .
- 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, 25th percentile = 5, 75th 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?

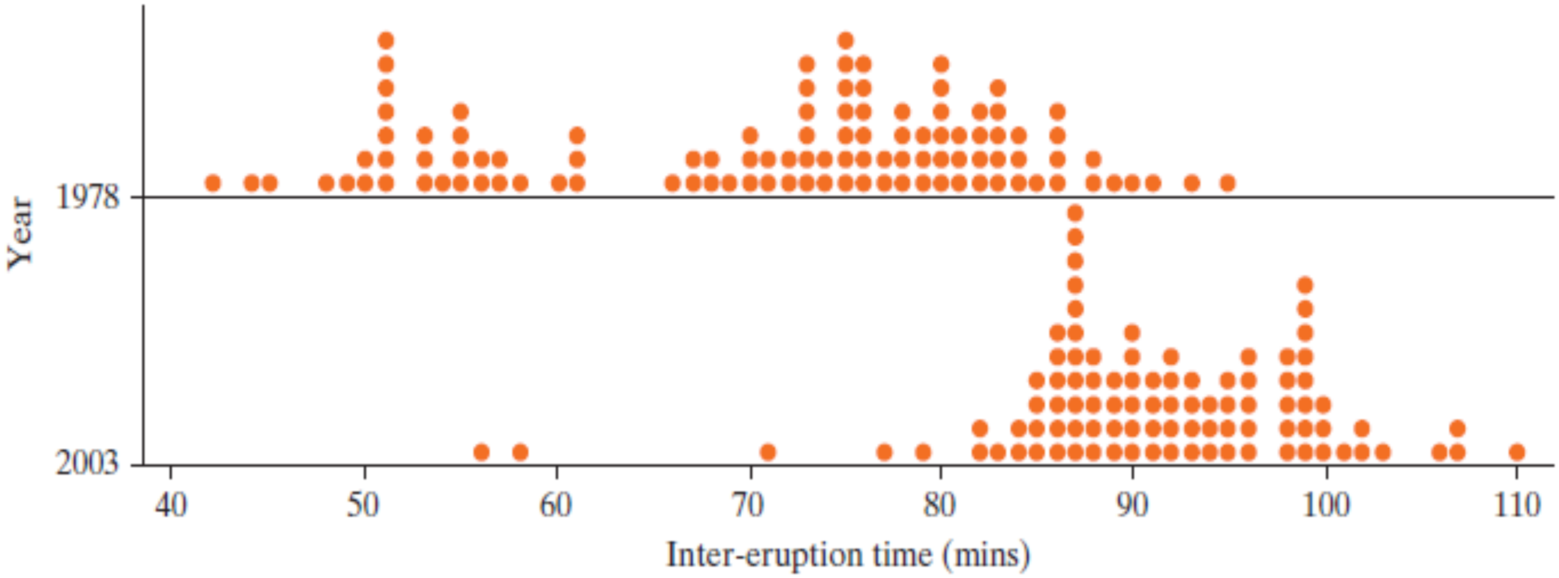
- For example, suppose data are 1, 3, 7, 7, 8, 9, 12, 14.
- M = 7.5, 25th percentile = 5, 75th percentile = 10.5. IQR = 5.5.

Geyser Eruptions

Example 6.1

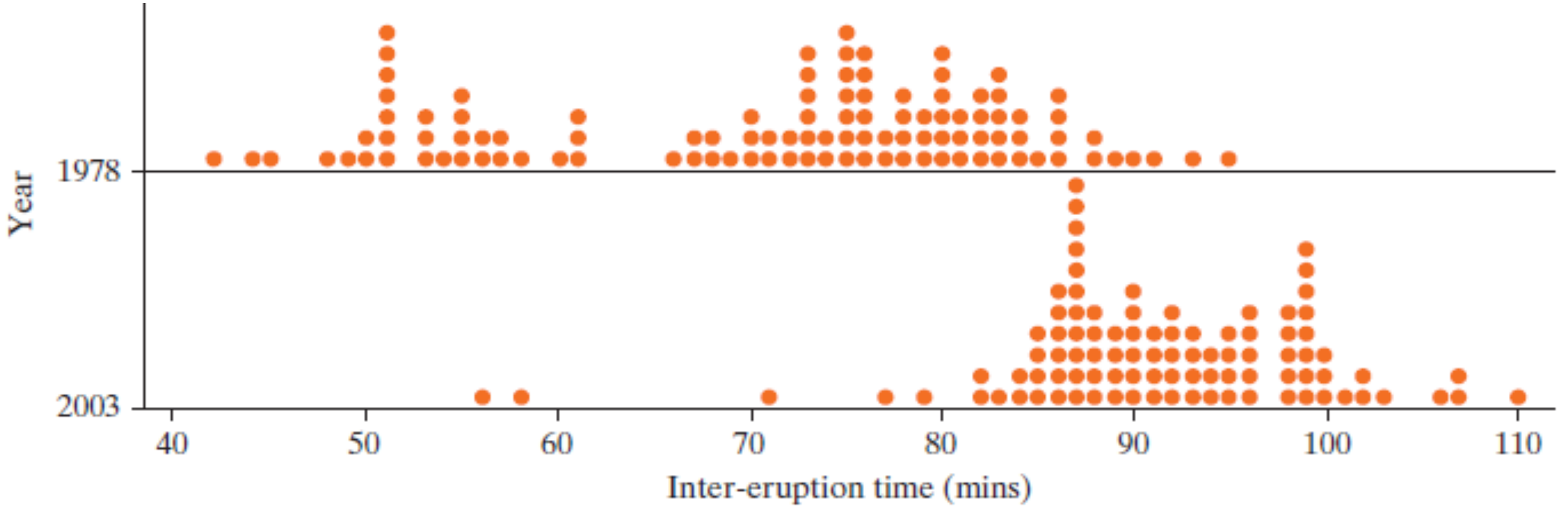
Old Faithful Inter-Eruption Times

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



Old Faithful Inter-Eruption Times

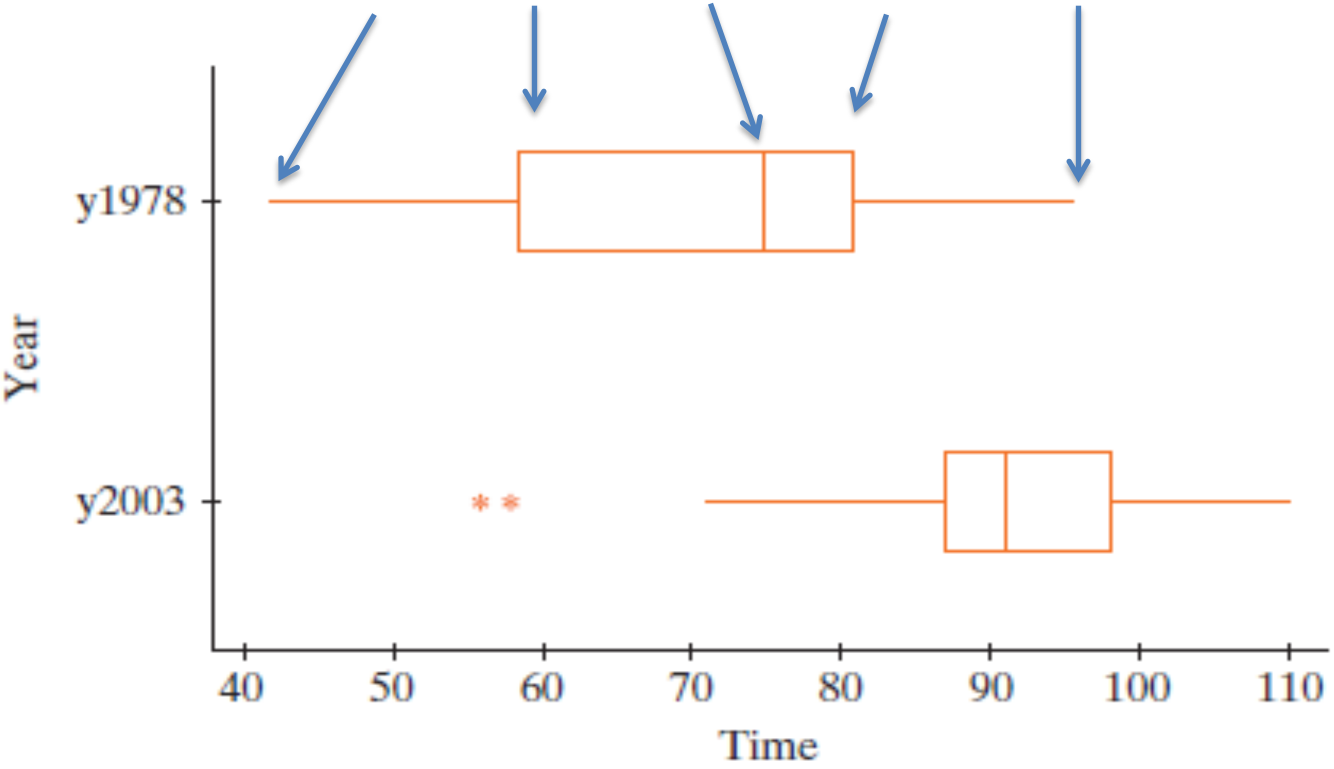
	Minimum	Lower quartile	Median	Upper quartile	Maximum
1978 times	42	58	75	81	95
2003 times	56	87	91	98	110



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

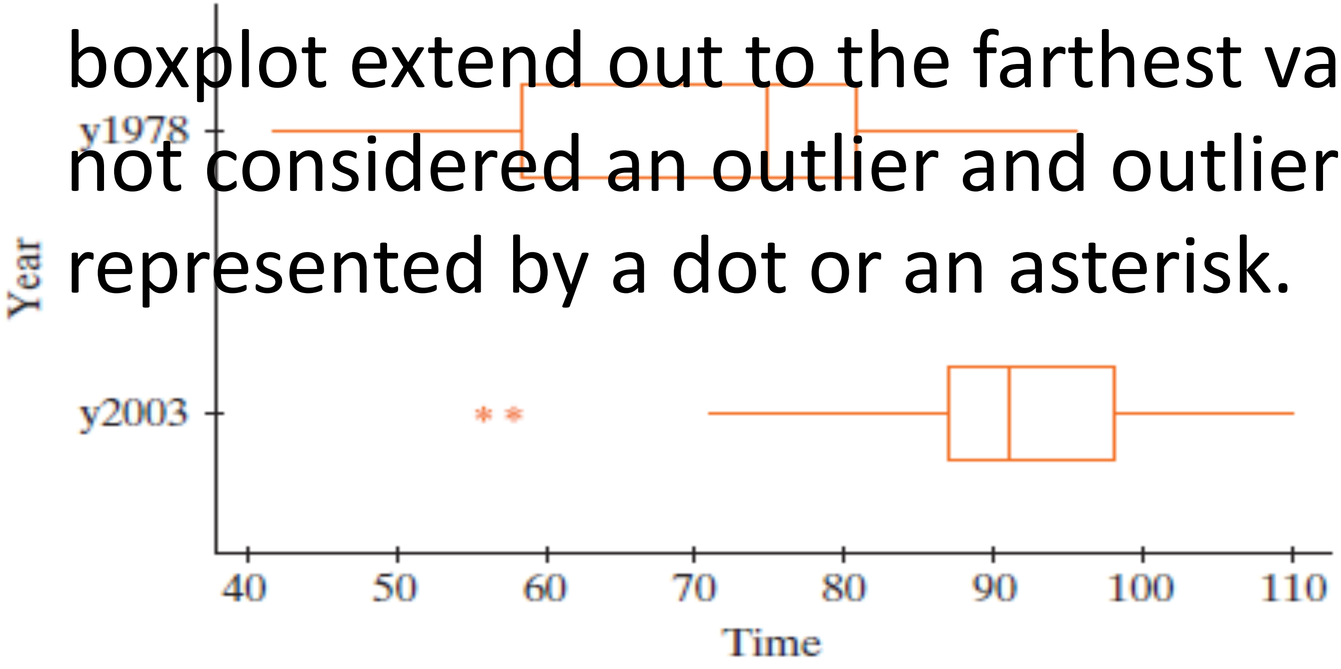
Boxplots

Min Q_{lower} Med Q_{upper} Max



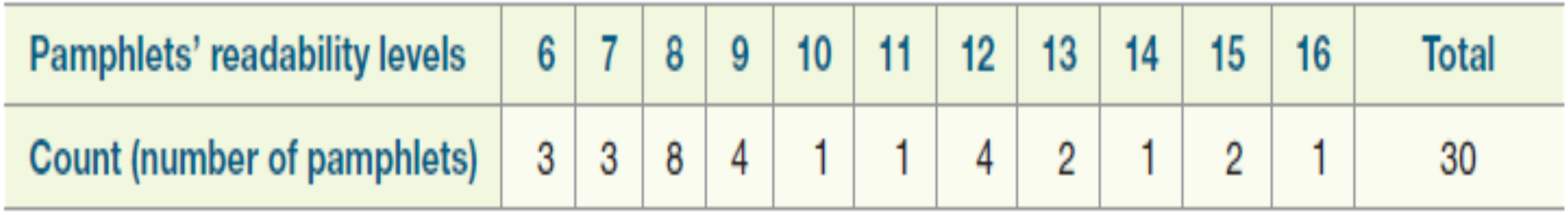
Boxplots (Outliers)

- 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.



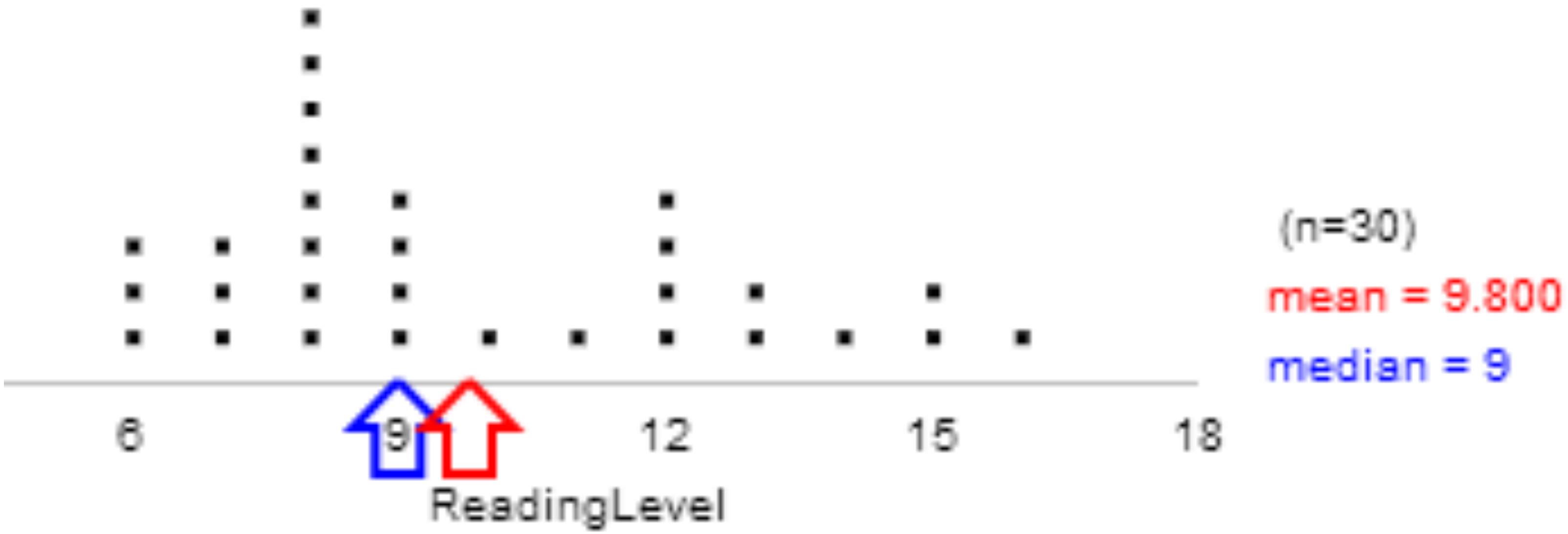
Cancer Pamphlet Reading Levels

- 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



2. 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.

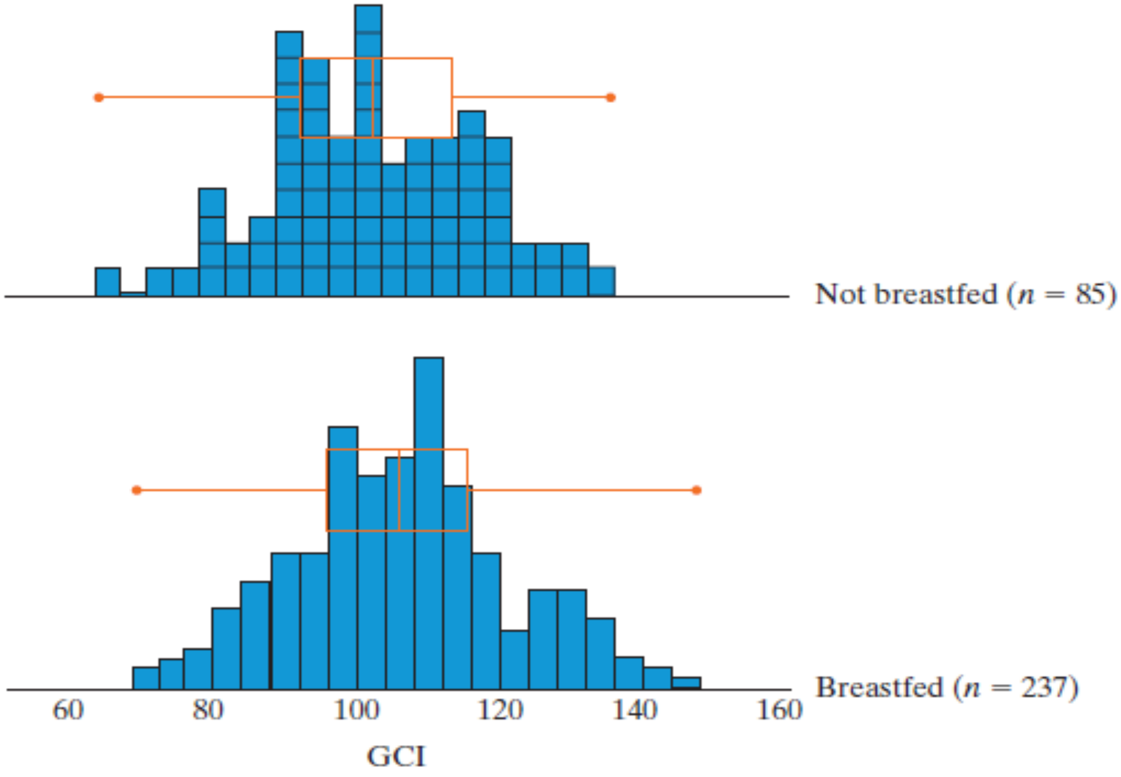
Breastfeeding and Intelligence

- $\mu_{breastfed}$ = 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.

3. 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.

4. 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 a Simple Random Sample (SRS) from the same large population.

standard means mean 0 and SD 1.

t versus normal and assumptions.

CLT: $(\bar{x}$ - μ) ÷ (σ/√n) -> standard normal.

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

So, if n is large, then

P($|(\bar{x}$ - μ) ÷ (σ/√n)| < 1.96) ~ 95%.

Mult. by (σ/√n) and get

P($|\bar{x}$ - μ| < 1.96 σ/√n) ~ 95%.

P($\mu - \bar{x}$ is in the range 0 +/- 1.96 σ/√n) ~ 95%.

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

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

t versus normal and assumptions.

CLT: $(\bar{x}$ - μ) ÷ (σ/√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, ..., X_n are iid and normal, then $(\bar{x}$ - μ) ÷ (σ/√n) is standard normal.

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

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

t versus normal and assumptions.

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

CLT: $(\bar{x}$ - μ) ÷ (σ/√n) ~ std. normal.

If X_1, ..., X_n are normal, then $(\bar{x}$ - μ)÷(σ/√n) is std. normal.

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

Gosset (1908) showed that replacing σ with s,

if X_1, ..., X_n are normal, then $(\bar{x}$ - μ)÷(s/√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 $\sigma/\sqrt{n}$) ~ 95%.

If the observations are iid and normal, then

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

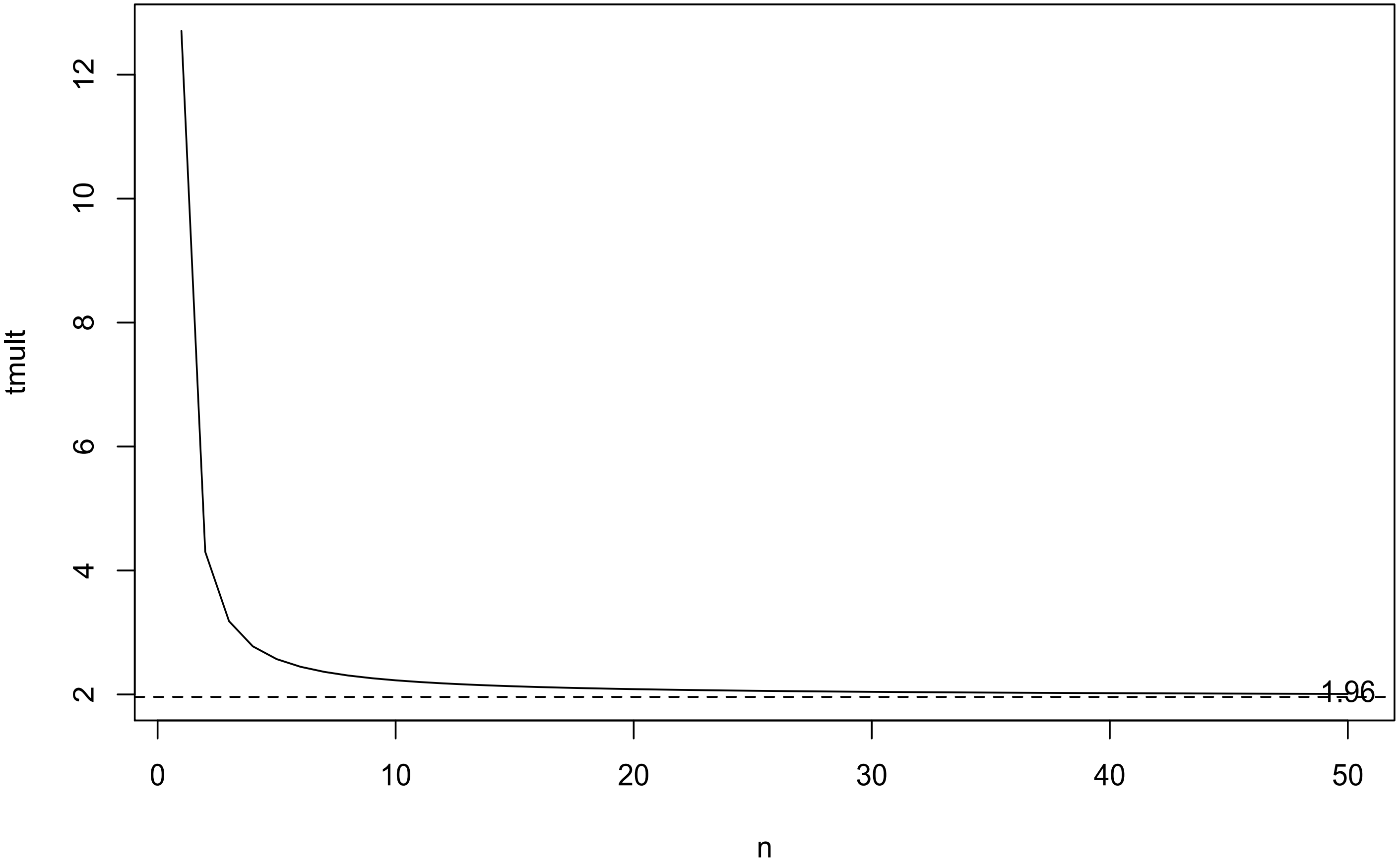
If the obs. are iid and normal and σ is unknown, then

P(μ is in the range $\bar{x}$ +/- t_{mult} $s/\sqrt{n}$) ~ 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 = √[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 σ/√n.
- If n is small, draws are normal, and σ is known, use $\bar{x}$ +/- 1.96 σ/√n.
- If n is small, draws ~ normal, and σ is unknown, use $\bar{x}$ +/- t_{mult} s/√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)], σ = √[π(1–π)].

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 μ_1 - μ_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 π_1 - π_2.

same as in c above, with $p_1 = \bar{x_1}$, s_1 = √[p_1 (1-p_1)], $\sigma_1 = \sqrt{}[\pi_1 (1-\pi_1)]$.

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

For testing, use pooled estimate of p for the SE.

For CIs for the difference in proportions,

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

In testing the difference in proportions,

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

where $\hat{p}$ is the proportion in both groups combined.