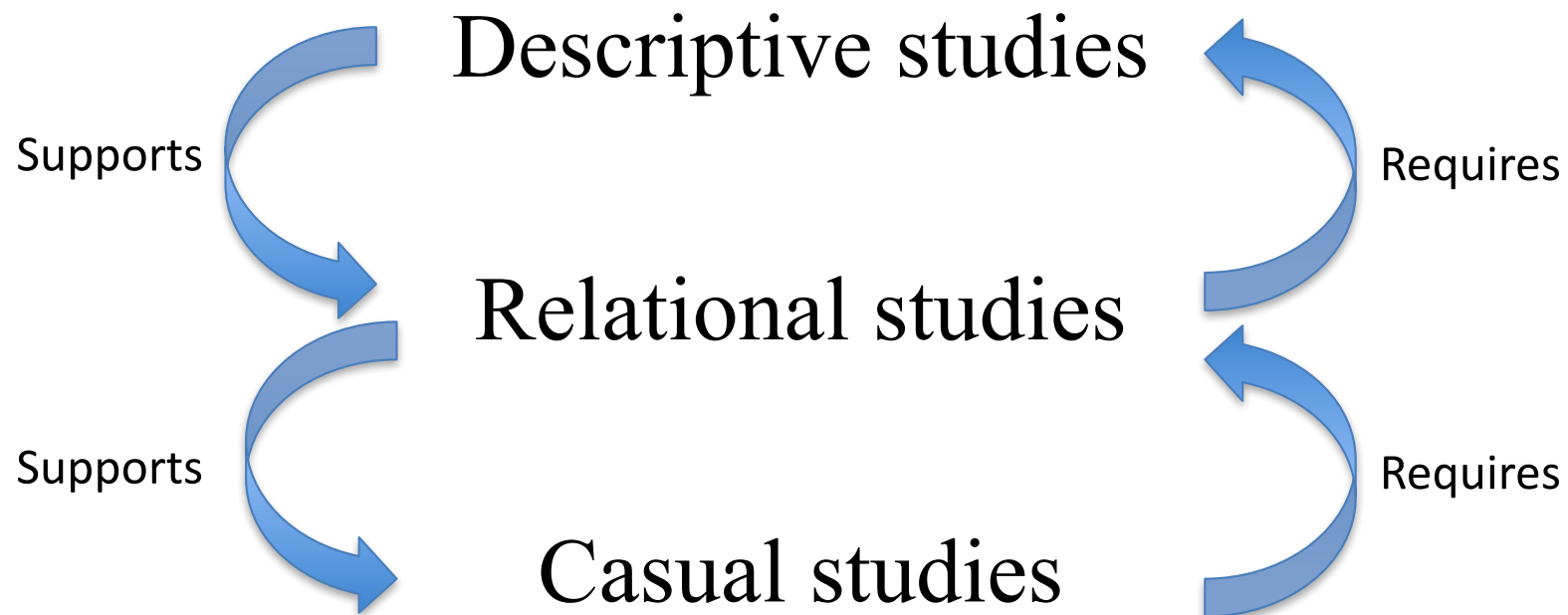
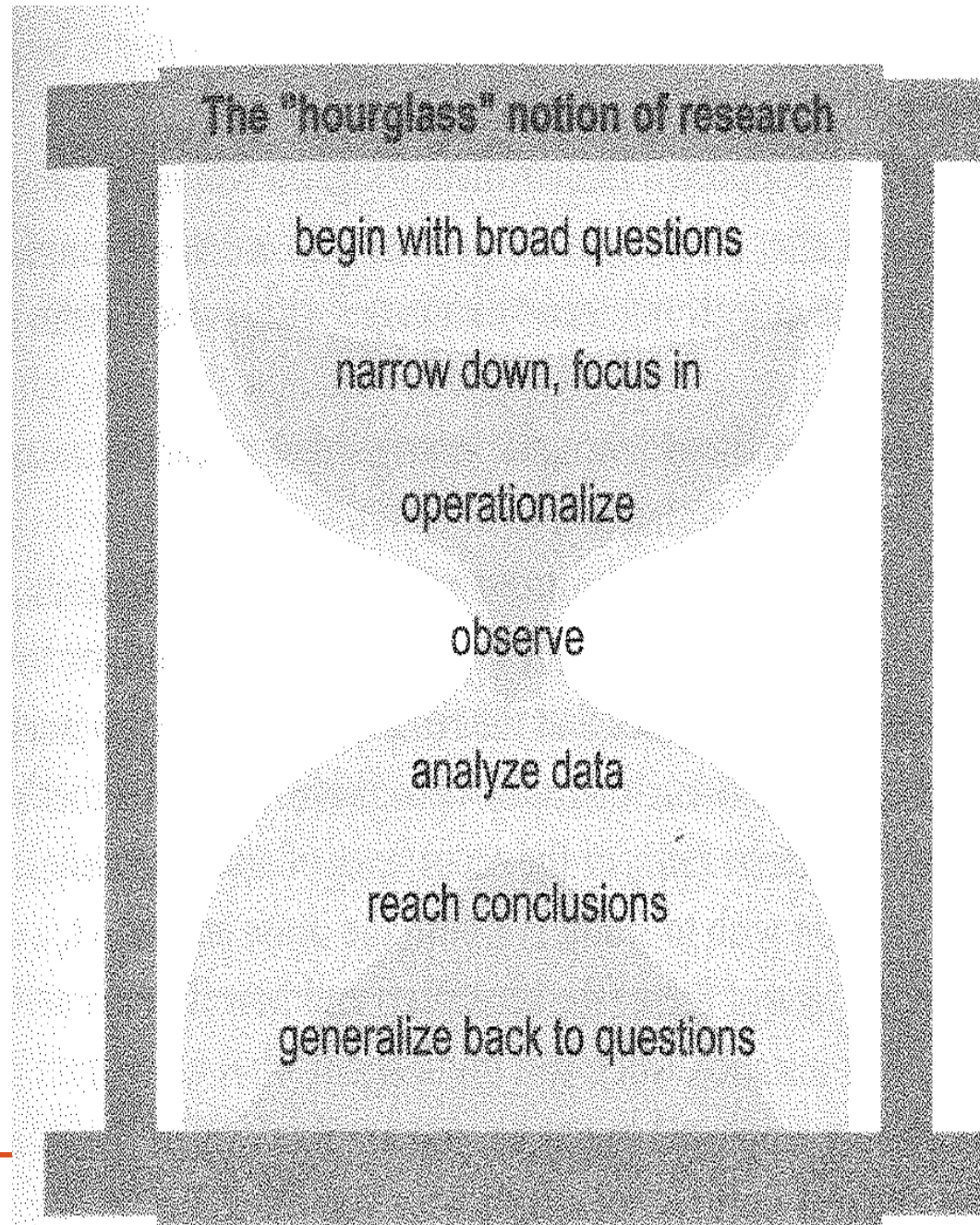


# Evaluation and Analysis

David M. Neyens, PhD MPH  
Assistant Professor  
Industrial Engineering  
[dneyens@clemson.edu](mailto:dneyens@clemson.edu)

# Types of studies





# Phases of Experimental Research

- Experimental or planning phase
- Design Phase
- Analysis Phase
- Interpretation Phase

# Experimental Phase

- Define problem statement
  - Do a literature review
  - Develop a hypothesis
- Define the variables of interest
  - Dependent variables: measures performance
    - Cannot be manipulated by experimenter (response variables)
    - Determine how measures should be scaled (nominal, ordinal, interval, ratio)
  - Independent variables (IV): controlled by experimenter
    - Controlled experimentally
    - Controlled statistically

# Experimental Phase

## Independent Variables (IV)

- Controlled by experimenter
- However, you need to determine whether the levels should be held constant or determined by a process of randomization
  - Fixed effects: factors whose levels are set at specified values (e.g., System A, B, or C)
  - Random effects: levels are chosen at random from among all possible levels (e.g., drivers)

# Experimental Phase

## Independent Variables (IV)

- Rigidly controlled: variables remain fixed throughout the experiment
  - e.g., The effects of 3 weight reducing programs (A, B, C) on weight loss.
  - The IV is the programs, the DV is weight loss
- Manipulated or set, at levels of interest: can be qualitative or quantitative; fixed or random
  - e.g., temperature, age
- Randomized: Order of experimentation should be randomized to average out the effects of variables that cannot be controlled (extraneous variables)

# Design Phase

- Size of the sample
- How many observations for each person
- How large a difference to be detected
- What variables can you control?
- What variables can you not control (e.g., weather)
- Randomize order of experimentation
- Set up a mathematical model to describe experiment



# Experiments

- Types of experiments
  - Completely randomized design (factorial experiment)
  - Repeated measures (each subject goes through multiple days, trials, etc.)
  - Restrictions on randomization
  - Correlated dependent variables (lane deviation, steering wheel position)
  - Unbalanced designs (each treatment group does not have an equal number of observations)
- Each situation requires the researcher to set up a different analysis (e.g., t-test, ANOVA (analysis of variance), regression model, or more complex models)

## Experimental Phase: Example

- Interested in determining the effects of drivers and type of in-vehicle system on driver speed. Researcher considers 3 in-vehicle devices and randomly chooses 5 drivers
- DV = average speed
- IV = in-vehicle device (fixed), drivers (random)

---

		Driver				
		1	2	3	4	5
Device	1					
	2					
	3					

---

Factorial Experiment: all levels are randomly assigned

# Design Phase: Mathematical Model

response variable =  $\mu + IV1_i + IV2_j + \dots$   
+ (interactions) + ... + (any restrictions imposed on the experiment)

From our example :

$$y_{ijk} = \mu + (\text{Device})_i + (\text{Operator})_j + D * O_{ij} + \varepsilon_{(ij)k}$$

where

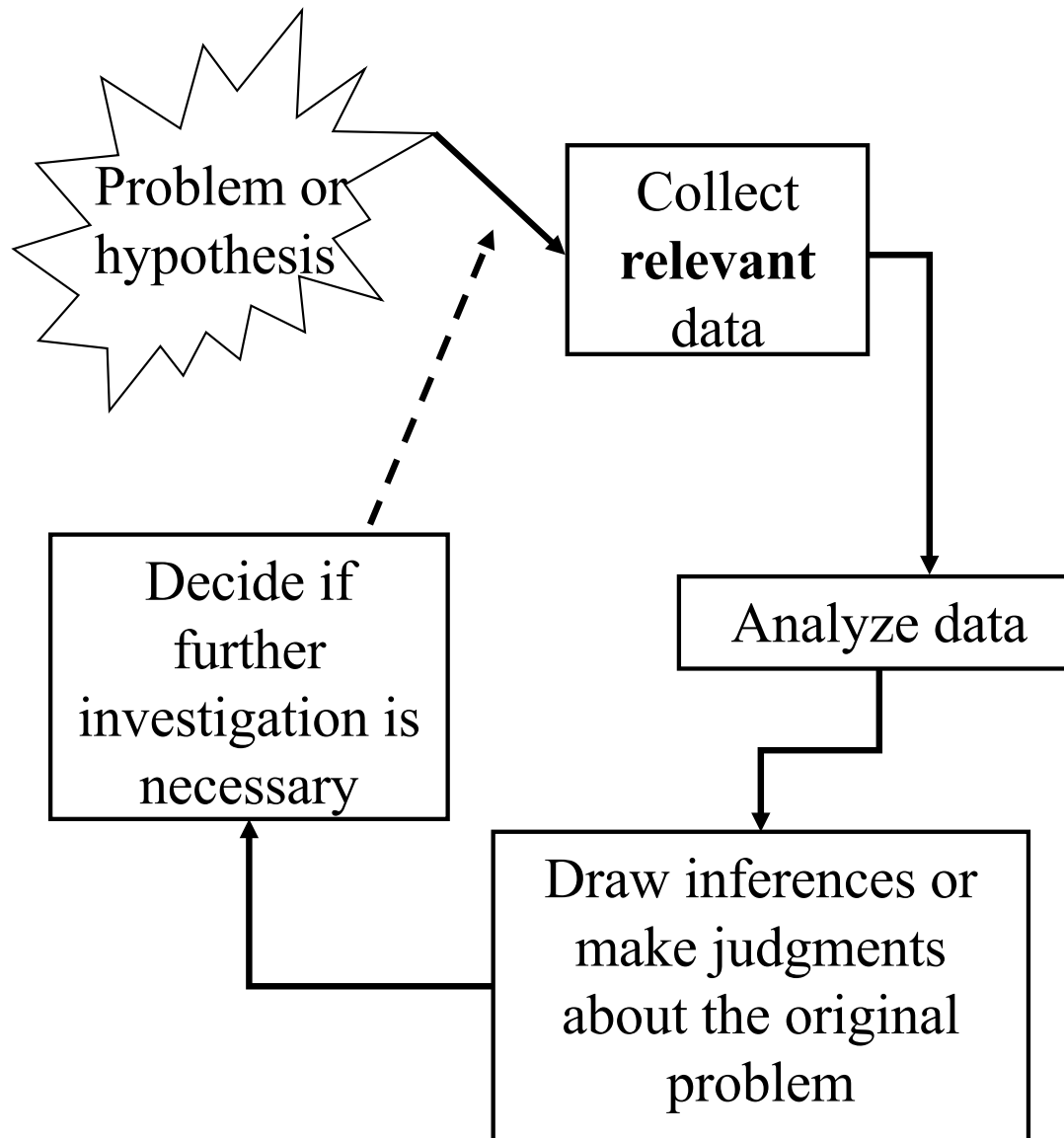
$$i = 1, 2, 3$$

$$j = 1, 2, 3, 4, 5$$

Formulate your null hypothesis

$$H_0: \text{Device}_1 = \text{Device}_2 = \text{Device}_3$$

# Hypothesis Testing



# Hypothesis Testing

- Hypothesis testing is a procedure to test a theory (statement) about parameters of one or more populations
- Some examples uses of hypothesis testing:
  - Does a new material have strength exceeding 200psi?
  - Is the variability of parts produced by a new process significantly lower than when using the current process?
  - Is the fraction of defective items from supplier 1 significantly lower than from supplier 2?
  - Is flu medicine 1 more effective than flu medicine 2?

# Hypothesis Testing Analogy

## Court room

Court trial →

Evidence is presented →

Cannot know if TRULY guilty

Assumed innocent until proven guilty

Defendant is guilty →

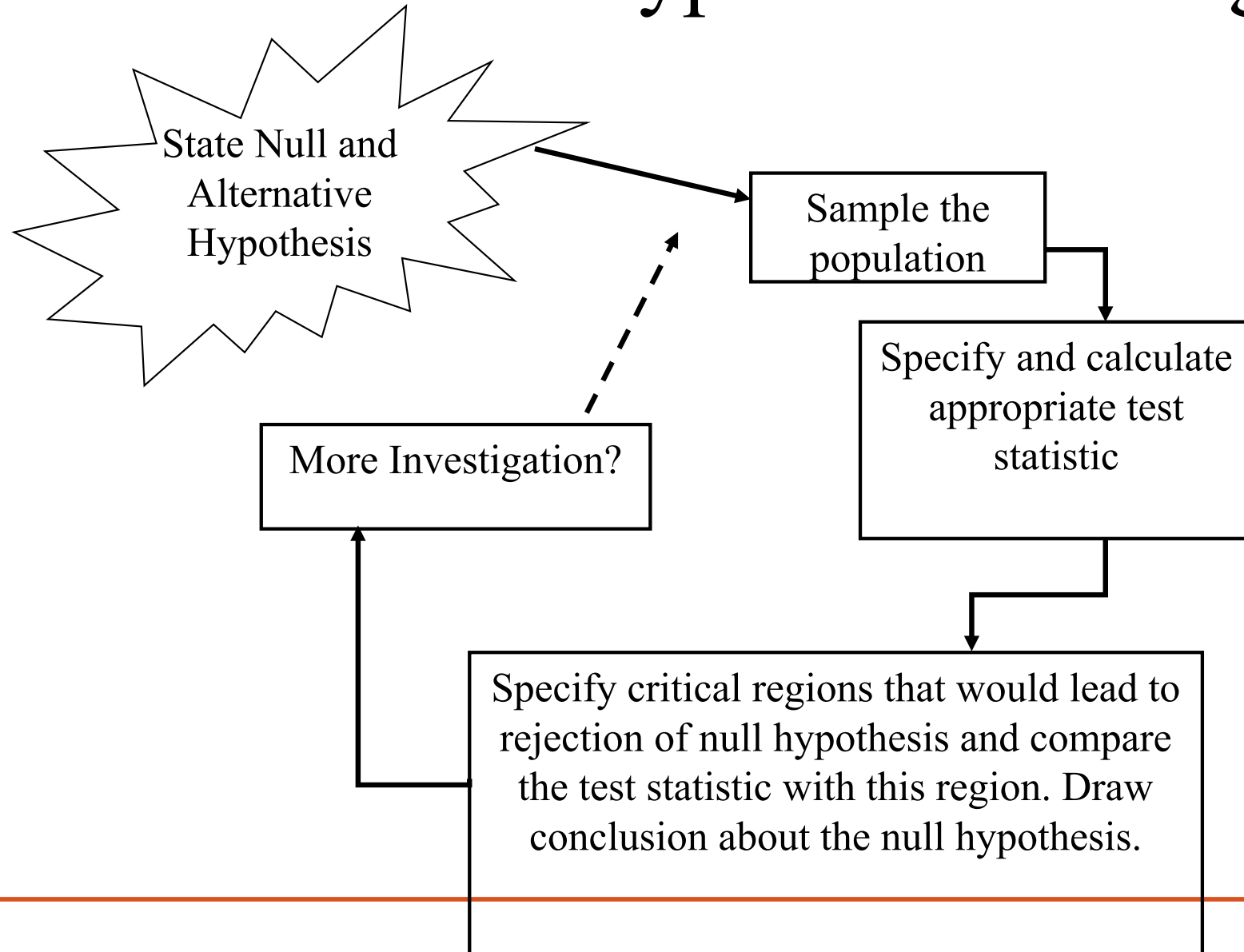
Defendant is not guilty  
(presumed innocence) →

Convict an innocent person →

Let guilty go free →

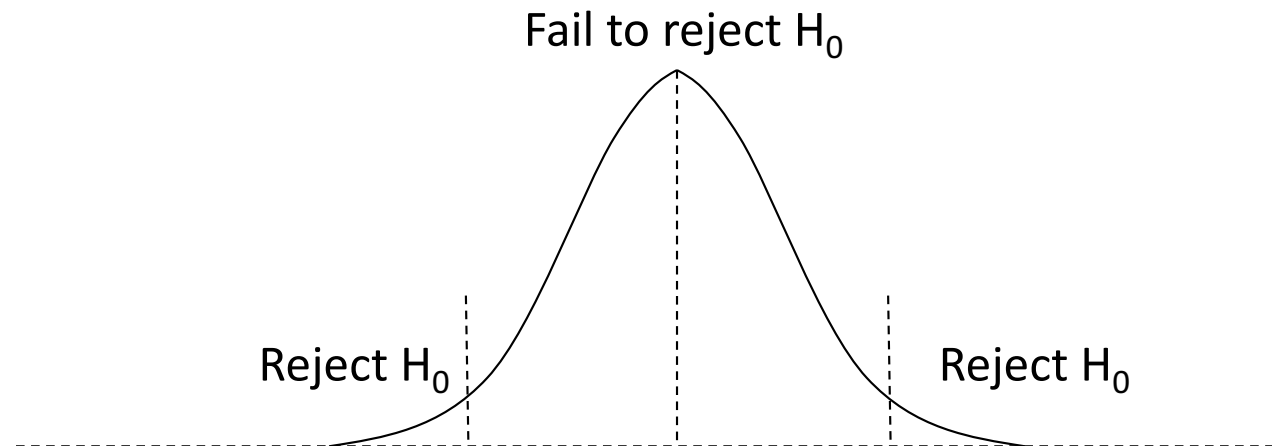
## Hypothesis testing

# Procedure for Hypothesis Testing



# Decisions in hypothesis testing

- If  $H_0: \mu=50$
- $H_A: \mu \neq 50$





# Errors in Hypothesis testing

Beta =  $P(\text{type II error})$   
Alpha =  $P(\text{Type I error})$

State of the world			
		$H_0$ false	$H_0$ true
Fail to reject $H_0$	Type II error	No error	
Reject $H_0$	No error	Type I error	

# Strong versus weak conclusions

- Rejecting  $H_0$  is a strong conclusion
  - We can control alpha (probability of Type 1 error)
  - We set it a priori to minimize the risk of Type 1 errors
- Failing to reject  $H_0$  is a weak conclusion.  
Why?

# Setting up null and alternative hypotheses

- One-sided hypothesis:

$$H_0: \mu = \mu_0$$

$$H_1: \mu < \mu_0$$

Reject  $H_0$  if  $Z_0 < -z_\alpha$

$$H_0: \mu = \mu_0$$

$$H_1: \mu > \mu_0$$

Reject  $H_0$  if  $Z_0 > z_\alpha$

- Two-sided

$$H_0: \mu = \mu_0$$

$$H_1: \mu \neq \mu_0$$

*Reject  $H_0$  if  $Z_0 < -z_{\alpha/2}$  or  $Z_0 > z_{\alpha/2}$*

## P-value in a hypothesis test

- P-value is NOT the probability that the null hypothesis is false and  $1-P$  is NOT the probability that the null hypothesis is true
- Interpreted as: the risk associated with wrongly rejecting the null hypothesis.

## Example problem

- An engineer is considering a new nickel-chrome-iron alloy. She has ordered 100 sample castings, which are to be tested at a materials lab for endurance under axial stress.
- She wants a metal strong enough to meet customer specifications for parts in a new stamping machine. This requires that the mean number of cycles until failure obtained in vibration testing should exceed 500,000. From specs, let  $\sigma=48,732$ .

## Procedure

1. Formulate the hypothesis. State the null hypothesis,  $H_0$  and the alternative hypothesis  $H_1$  about some parameter  $\theta$ .
2. Select the test statistic.
3. Establish the significance level and the acceptance and rejection regions for the decision rule.
4. Compute the value of the test statistic from the data.
5. Make the decision.

# Comments

- Test aimed to assess strength of evidence against null hypothesis. Either we have enough evidence to reject the null hypothesis or we do not have enough evidence to reject.
- The null hypothesis is the theory we hope to reject, the alternative hypothesis is what we want to support, by rejecting the null hypothesis.
- We can only reject hypotheses. We can never prove a hypothesis, all we can say is, we had insufficient evidence to reject the hypothesis: The **strong** conclusion is to provide sufficient evidence to reject the hypothesis.

# Research fallacies and validity



# Types of fallacies

- Fallacies → Errors in reasoning, usually based on assumptions
  - Ecological fallacy → assuming single observation is part of average
  - Exception fallacy → assume single observation implies averages

# Validity

- Validity applied to propositions, inference, or conclusions.
- There are 4 main types of validity (but some have subcomponents)
  - Conclusion
  - Internal
  - Constructs
  - External

# Why is validity important?

- To accurately interpret cause-effect relationships
- Conclude that the manipulation influences outcome measures
- Is the evidence good or poor?
- If there is low validity in a study, then we should not draw conclusions from that study.
  - Thus, without validity a study is garbage.

# Conclusion validity

- Is there a relationship between the variables?
  - What are some threats to conclusion validity?

# Internal validity

- If there is a relationship between the variables (conclusion validity) then is the relationship causal?
  - What are some threats to internal validity?  
Hint: There are LOTS!

# Establishing causality

- Temporal precedence → Timing matters
- Covariation of the cause and effect →  
Relationship between variables
- No plausible alternative explanations →  
Cannot contribute effect to extraneous variable

# Threats to Internal Validity

- Single-group threats
  - Only a single group receives intervention
- Multiple-group threats
  - When several groups are included  
(differences/similarities between groups)
- Social threats to internal validity
  - When research is ‘less controlled’ there may be other factors that influence behavior

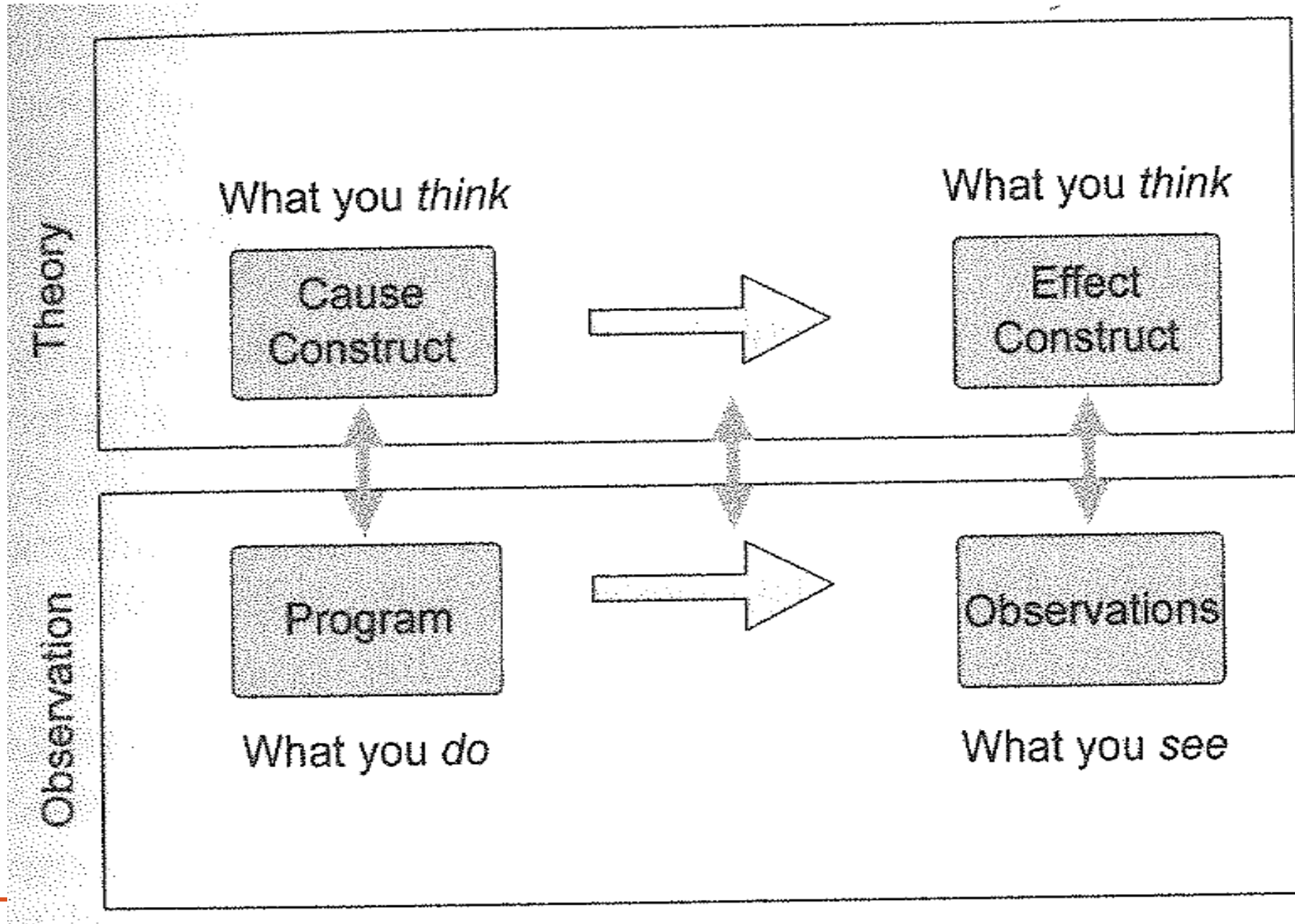
# Threats to Internal Validity

- History threat
- Maturation threat
- Testing threat
- Instrumentation threat
- Mortality threat (study drop outs)
- Regression threat (regression to the mean, non-random sample)



# Construct Validity

- Assuming a causal relationship, does the relationship support the constructs of the intervention and the construct of the measurement
  - It is the intervention what you intended and are you measuring what you intended.



# Within construct validity

- Translation validity
  - Face validity
  - Content validity
- Criterion-related validity
  - Predictive validity
  - Concurrent validity
  - Convergent validity
  - Discriminant validity

# Translational Validity

- Is the operationalization a good reflection of the construct?
- Face validity → On the surface does it support the construct?
  - Weakest way to demonstrate construct validity
- Content validity → How does the operationalization match the literature and domain expertise

# Criterion-related Validity

- Does the operationalization behaves the way it should in your theory?
- Predictive→ Can the operationalization correctly predict what it should be able to predict?
- Concurrent→ Can the operationalization correctly distinguish between groups?
- Convergent→ Is the operationalization similar to other operationalizations it should be similar to?
- Discriminant→ Is the operationalization different than operationalizations for which it should differ?

# Coke example



- For market researchers, **criterion validity** is crucial, and can make or break a product. One famous example is when Coca-Cola decided to change the flavor of their trademark drink.
- Diligently, they researched whether people liked the new flavor, performing taste tests and giving out questionnaires. People loved the new flavor, so Coca-Cola rushed New Coke into production, **where it was a titanic flop.**

# External validity

- Can the results of this study be generalized to other populations, groups, and situations?
  - What are threats to external validity?



# The Validity Questions are cumulative...

Validity

External

Can we generalize to other persons, places, times?

Construct

Can we generalize to the constructs?

Internal

Is the relationship causal?

Conclusion

Is there a relationship between the cause and effect?



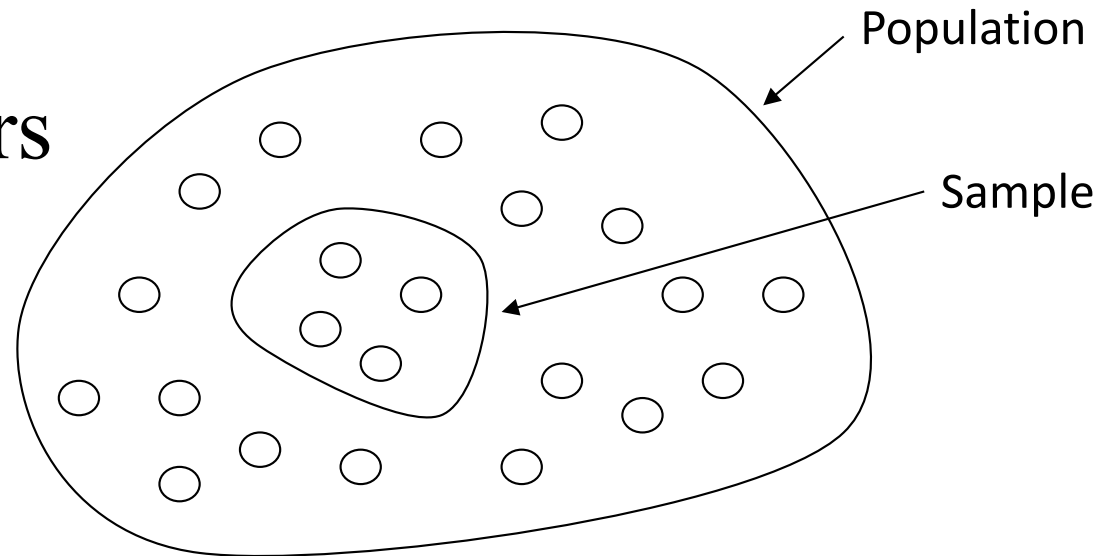
# How to minimize threats to validity

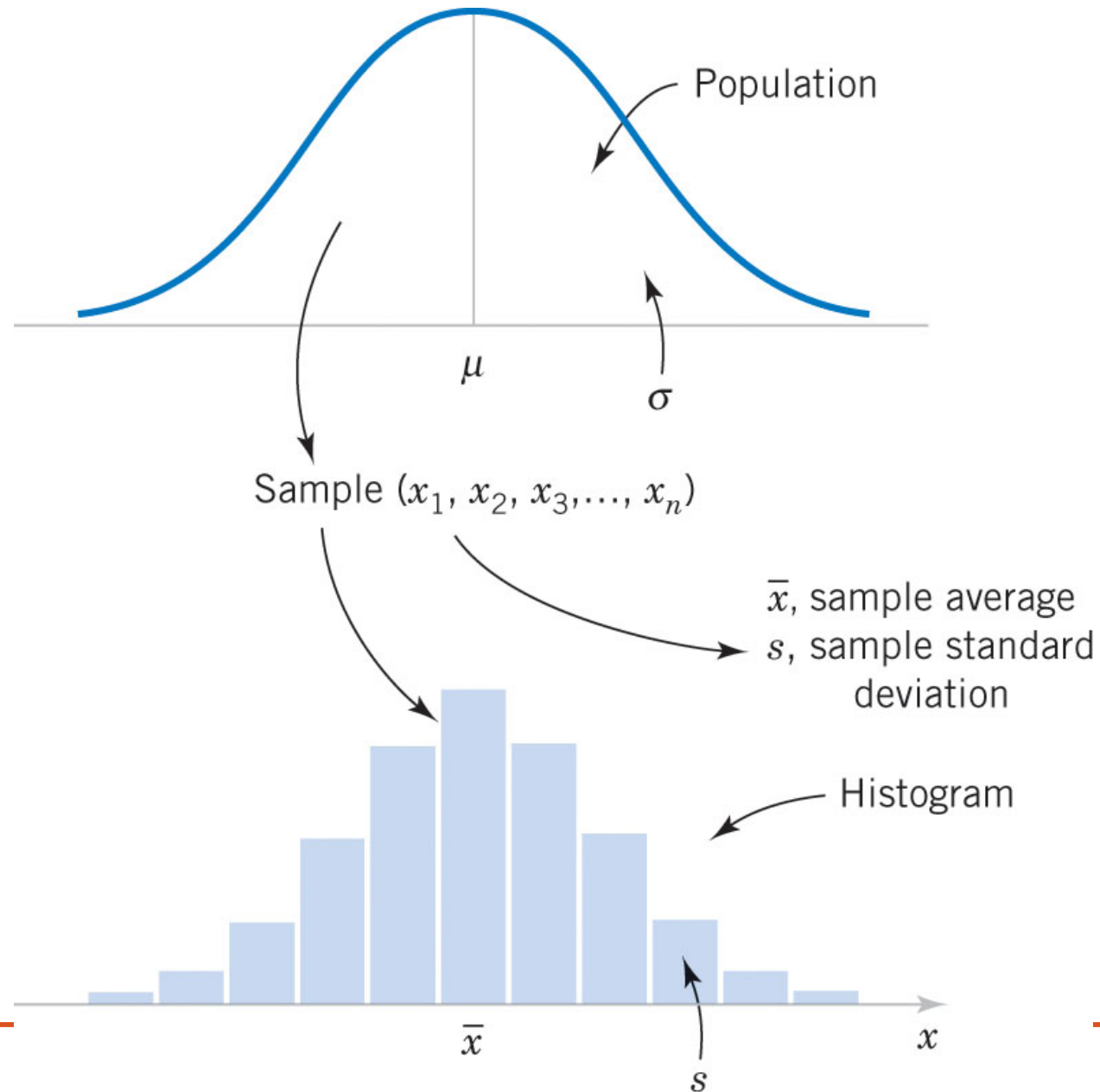
- By argument → Argue that the threat is not an issue
- By measurement or observation → If you can quantify some threat it may minimize the effect
- By design → Use appropriate control groups and designs
- By analysis → Provide an analysis that minimizes threats (e.g., regression on pre-test scores)

# Data analysis

# Sample from a population

- Typically in an engineering analysis, only a sample from a population is collected
- Population parameters
  - $\mu$ ,  $\sigma^2$
- Sample parameters
- $\bar{X}$ ,  $s^2$





# Numerical summaries of data

- Mean:            Population            Sample

$$\mu = \frac{\sum_{i=0}^N x_i}{N}$$

$$\bar{x} = \frac{\sum_{i=0}^n x_i}{n}$$

- Median  $\tilde{x}$

Separates the upper and lower 50% of the sample

Either the middle point if n=odd or the average of the two middle points if n=even

- Mode

The most frequently observed data point in the sample  
(not always a single value)

# Variance

Population

Sample

- Variance

$$\sigma^2 = \frac{\sum_{i=1}^N (x_i - \bar{x})^2}{N}$$

$$s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}$$

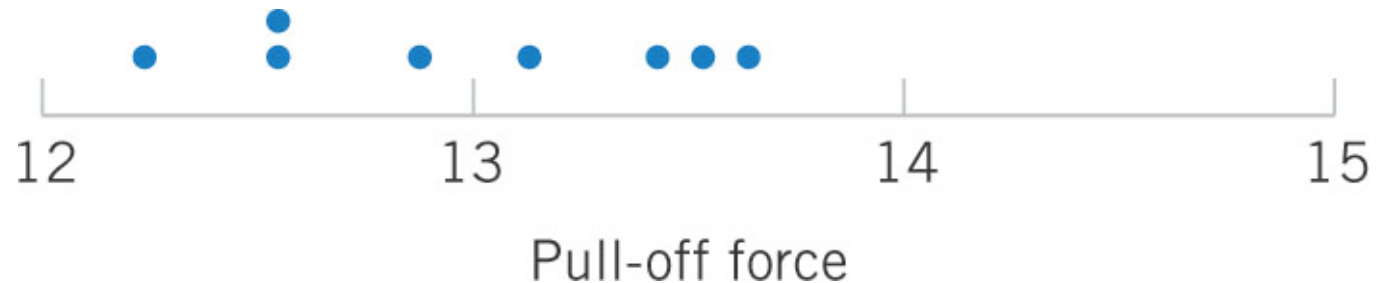
- Standard deviation

$$\sigma = \sqrt{\sigma^2}$$

$$s = \sqrt{s^2}$$

# Stem and leaf diagrams

- Dot diagram



- A stem and leaf diagram can be used with more data.
- Steps:
  - Divide each data point into two parts, a stem (typically the leading digit or digits) and the leaf (the remaining digits)
  - List all of the values in a vertical column
  - Attach each data point (leaf) to the stems
  - Document the units.

# Example of a stem and leaf plot

- What is the range of the data?

- What is the mode?

- What is the median?

Stem	Leaf	Frequency
7	6	1
8	7	1
9	7	1
10	5 1	2
11	5 8 0	3
12	1 0 3	3
13	4 1 3 5 3 5	6
14	2 9 5 8 3 1 6 9	8
15	4 7 1 3 4 0 8 8 6 8 0 8	12
16	3 0 7 3 0 5 0 8 7 9	10
17	8 5 4 4 1 6 2 1 0 6	10
18	0 3 6 1 4 1 0	7
19	9 6 0 9 3 4	6
20	7 1 0 8	4
21	8	1
22	1 8 9	3
23	7	1
24	5	1



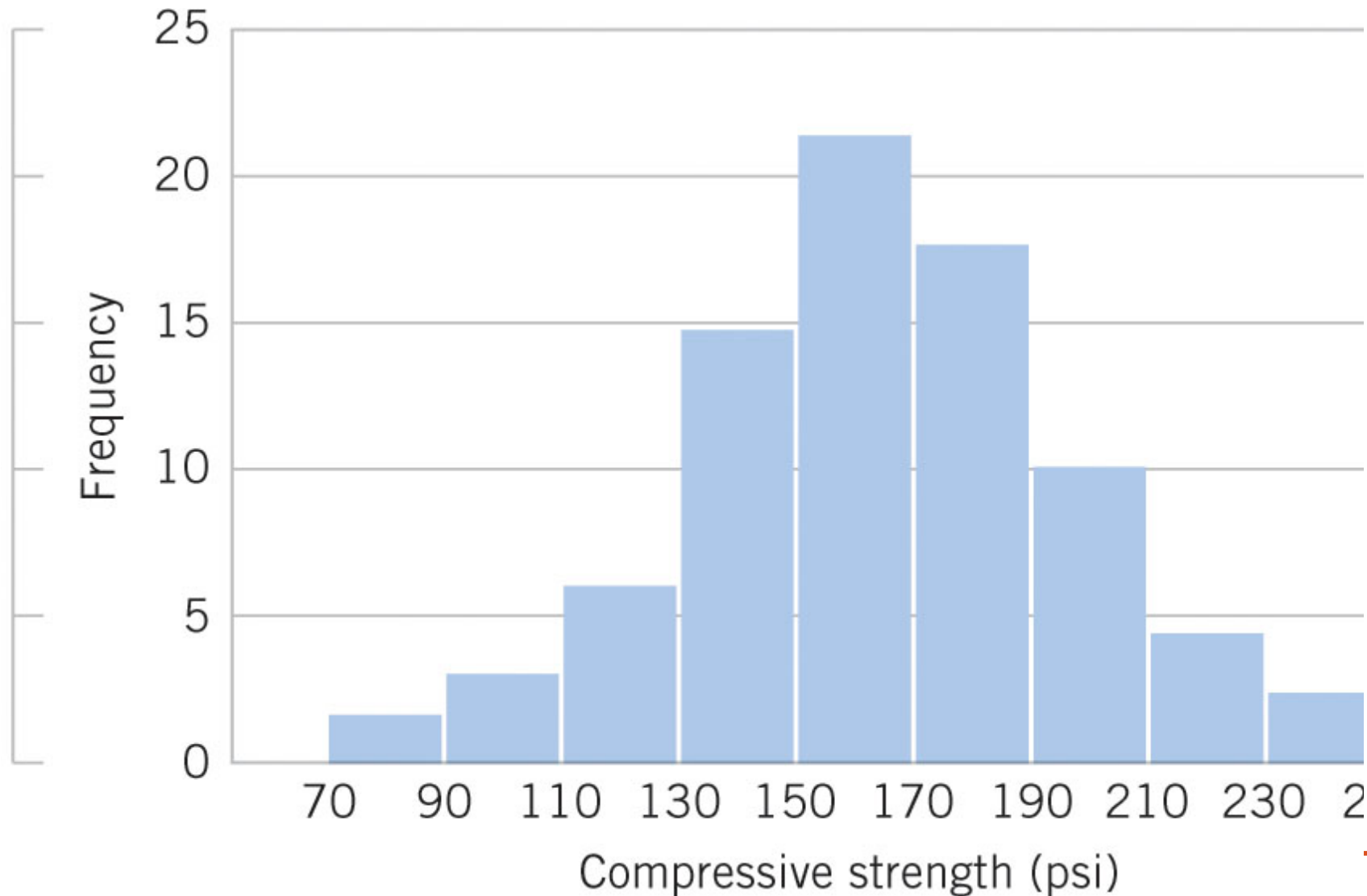
# Frequency distributions and histograms

- Bins are selected such that between 5 and 20 bins are used.

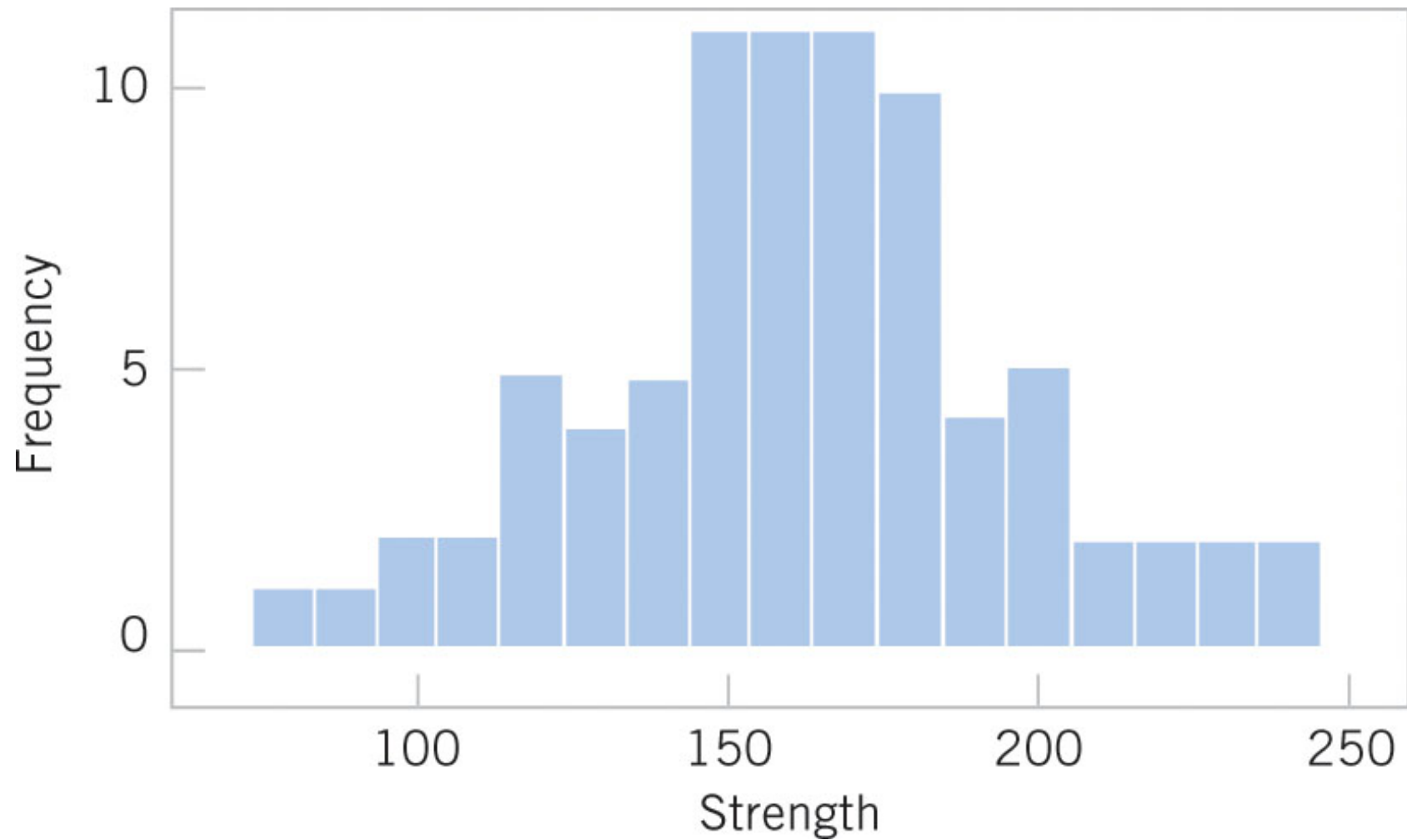
The number of bins is roughly  $\sqrt{n}$

- This is subjective, but needs to demonstrate the shape of the data.
- If the bins correspond to categories of data, not bins, then it is referred to as a Pareto chart.

9 bins, width=20

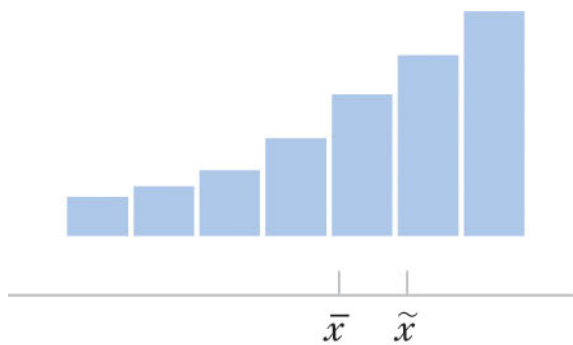


17 bins, width=10

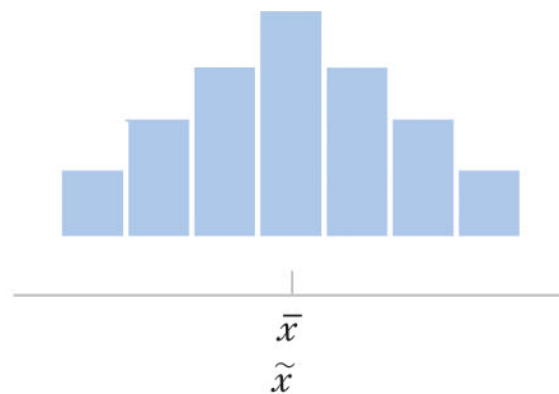


# Skewed distributions

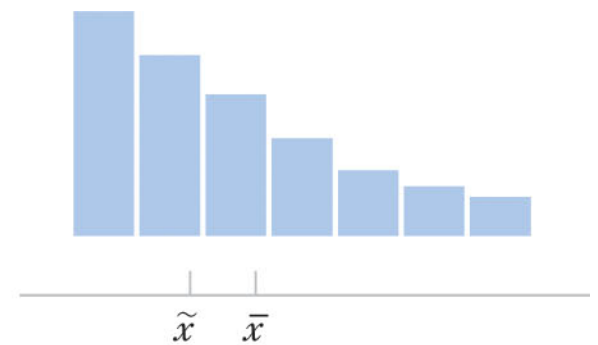
- Can identify if the distribution is skewed.
- Skewed left (long tail to the left), skewed right (long tail to right)



Negative or left skew  
(a)



Symmetric  
(b)



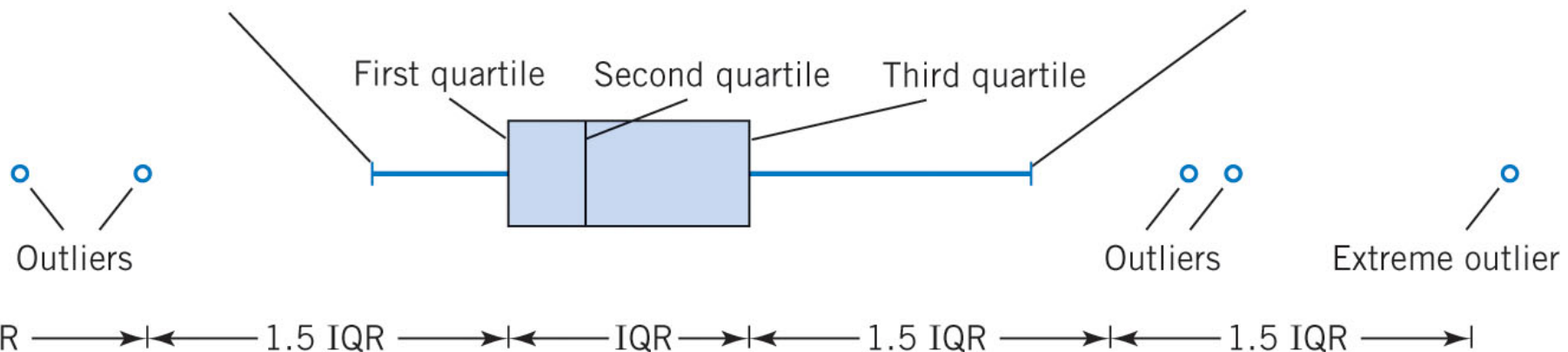
Positive or right skew  
(c)

# Boxplots (AKA box & whisker plot)

- Displays the quartiles (1<sup>st</sup>, 2<sup>nd</sup>, and 3<sup>rd</sup>)
- How do you calculate the quartiles?

Whisker extends to  
smallest data point within  
1.5 interquartile ranges from  
first quartile

Whisker extends to  
largest data point within  
1.5 interquartile ranges from  
third quartile



- Whiskers extend from 1<sup>st</sup> and 3<sup>rd</sup> quartiles out to the last data point that is within 1.5 times the IQR.
  - Does **not** extend the full  $1.5 \times \text{IQR}$  on both sides, only to the last data point on each side.



TL;DR

January 8 · 🌐

No. In fact there hasn't been any conclusive reports of anyone ever reporting to have swallowed a spider in their sleep.

"But a "fun fact" page on Facebook/Twitter/Instagram said we swallow 8 spiders every year in our sleep."

A false fact planted by mosquitos hoping people would subconsciously eat spiders in their sleep, thus allowing mosquito numbers to thrive.



Do we really swallow spiders in our sleep?

ANIMALS.HOWSTUFFWORKS.COM

👍 Like

💬 Comment

➦ Share

👍 😂 🍷 189

Top Comments ▾



**Fury Buri** "Average person eats 3 spiders a year" factoid actualy just statistical error. Average person eats 0 spiders per year. Spiders Greg, who lives in cave & eats over 10,000 each day, is an outlier adn should not have been counted

Like · Reply · 🗨️ 8 · January 8 at 10:22pm

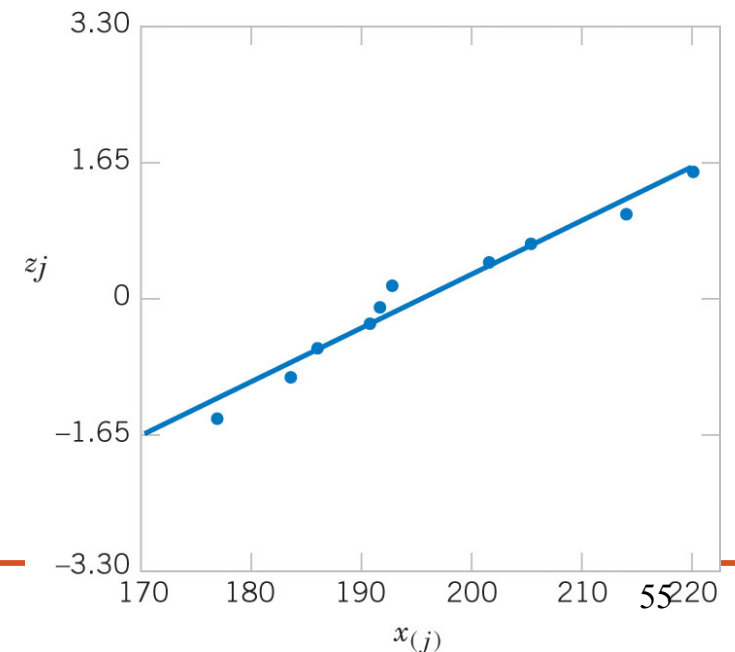


TL;DR Fucking Greg

Like · Reply · 🗨️ 5 · January 8 at 11:30pm

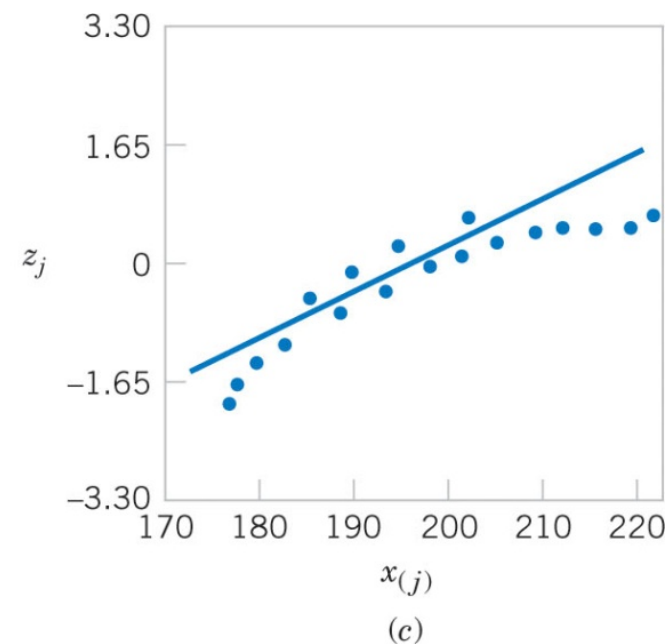
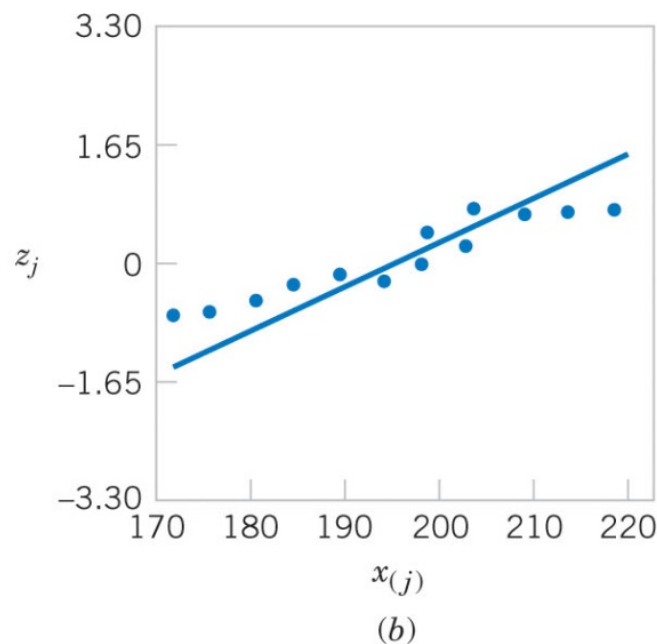
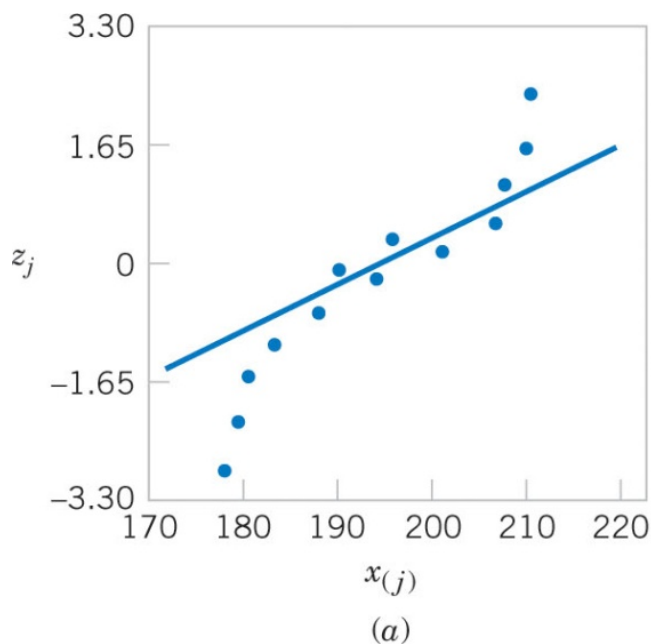
# Probability plots

- Used to determine if some data come from a specific probability distribution.
- X-Y plot
  - X axis is the values of the data
  - Y axis is usually the standardized normal scores for each value.



# Using normal probability plot

- Use “fat pencil” test to determine if the plot is approximately normal.





# Exercises

- When will the median of a sample be equal to the mean?
- When will the mode of a sample be equal to the median?

## Normal distribution

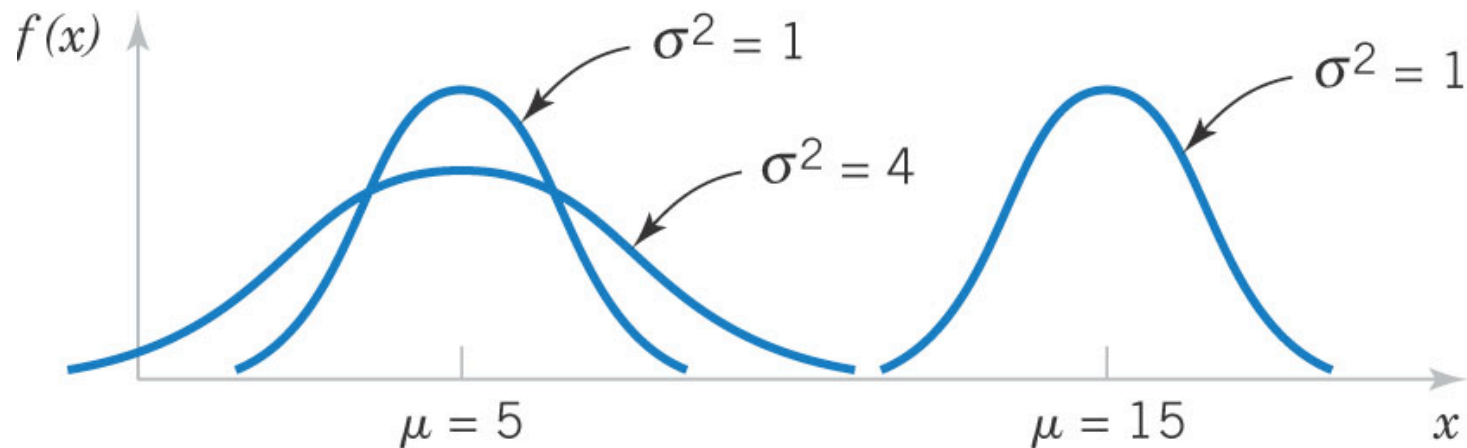
- Examples of data that is normally distributed?
- A continuous random variable  $X$  has a normal distribution with parameters  $\mu$  and  $\sigma$  (or  $\sigma^2$ ) written  $X \sim N(\mu, \sigma^2)$  where  $-\infty < \mu < +\infty$  and  $0 < \sigma < +\infty$ , if the probability density function is:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}} \quad \text{for} \quad -\infty < x < \infty$$

- $E[X] = \mu$
- $V[X] = \sigma^2$

# Distribution characteristics

## Bell shaped distribution



Symmetry of  $f(x)$ :  $P(X < \mu) + P(X > \mu) = 1.0$

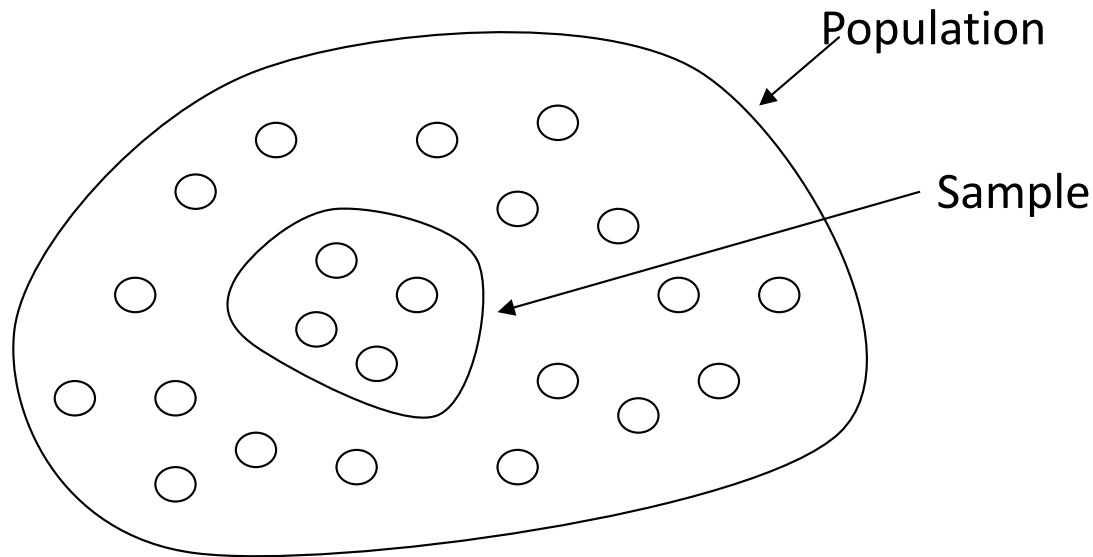
$P(\mu - \sigma < X < \mu + \sigma) = 0.6827$

$P(\mu - 2\sigma < X < \mu + 2\sigma) = 0.9545$

$P(\mu - 3\sigma < X < \mu + 3\sigma) = 0.9973$

# Statistical Inference

- Draw a sample from a population and infer from the sample about the whole population.
- Take a random sample  $X_1, X_2, \dots, X_n$  is of size  $n$



# Point Estimation

- A point estimate is a reasonable value of a population parameter
- Often we don't know the true probability distribution for a population. We sample the population to make inferences on the population's probability distribution.
- Distributions are described by their parameters.
- For example, what parameters describe the
  - normal distribution?
  - binomial distribution?

# Point Estimation

- A point estimate of a **parameter**  $\theta$  is obtained by selecting an appropriate **statistic**  $\Theta$ , and computing its value using the sample data. The selected statistic is called the point estimator.

# Statistical inference

- Making decision about a population based on information from a random sample from that population
- A RANDOM sample is very important.
  - Must be independent random variables
  - All data must have the same probability distribution

- A **statistic** is any function of the observations in a random sample
- Sampling distribution: The probability distribution of a statistic

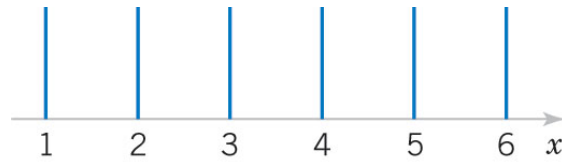


# Central Limit Theorem

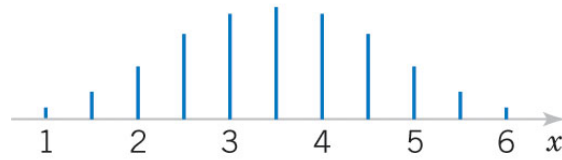
- Let  $X_1, X_2, \dots, X_n$  be a random sample of size  $n$  taken from a population with mean  $\mu$  and finite variance  $\sigma^2$ . If  $\bar{X}$  is the sample mean, then the limiting form of the distribution of

$$Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$$

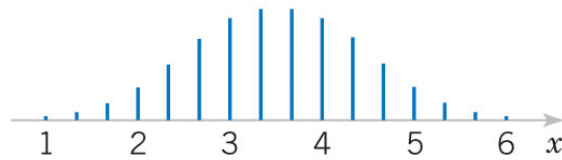
as  $n \rightarrow \infty$  is the standard normal distribution.



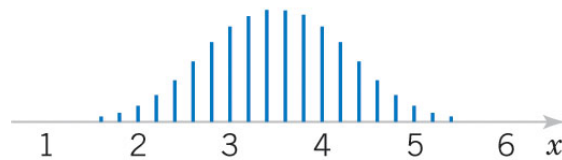
(a) One die



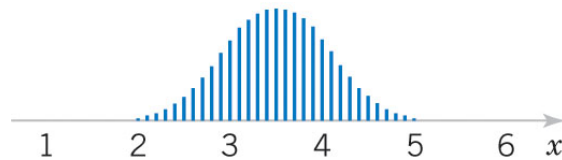
(b) Two dice



(c) Three dice



(d) Five dice



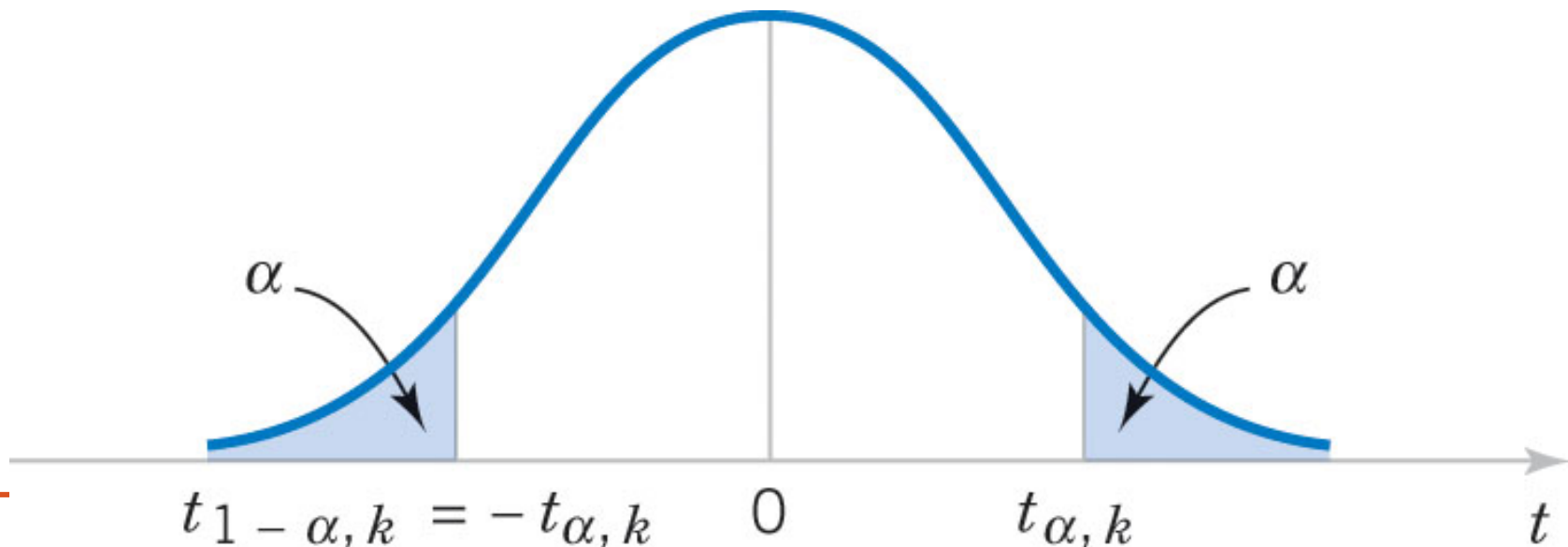
(e) Ten dice

# Intervals

- We can generate an interval for an estimate for a parameter
- Confidence intervals
- Tolerance intervals
- Prediction intervals

# Confidence Interval

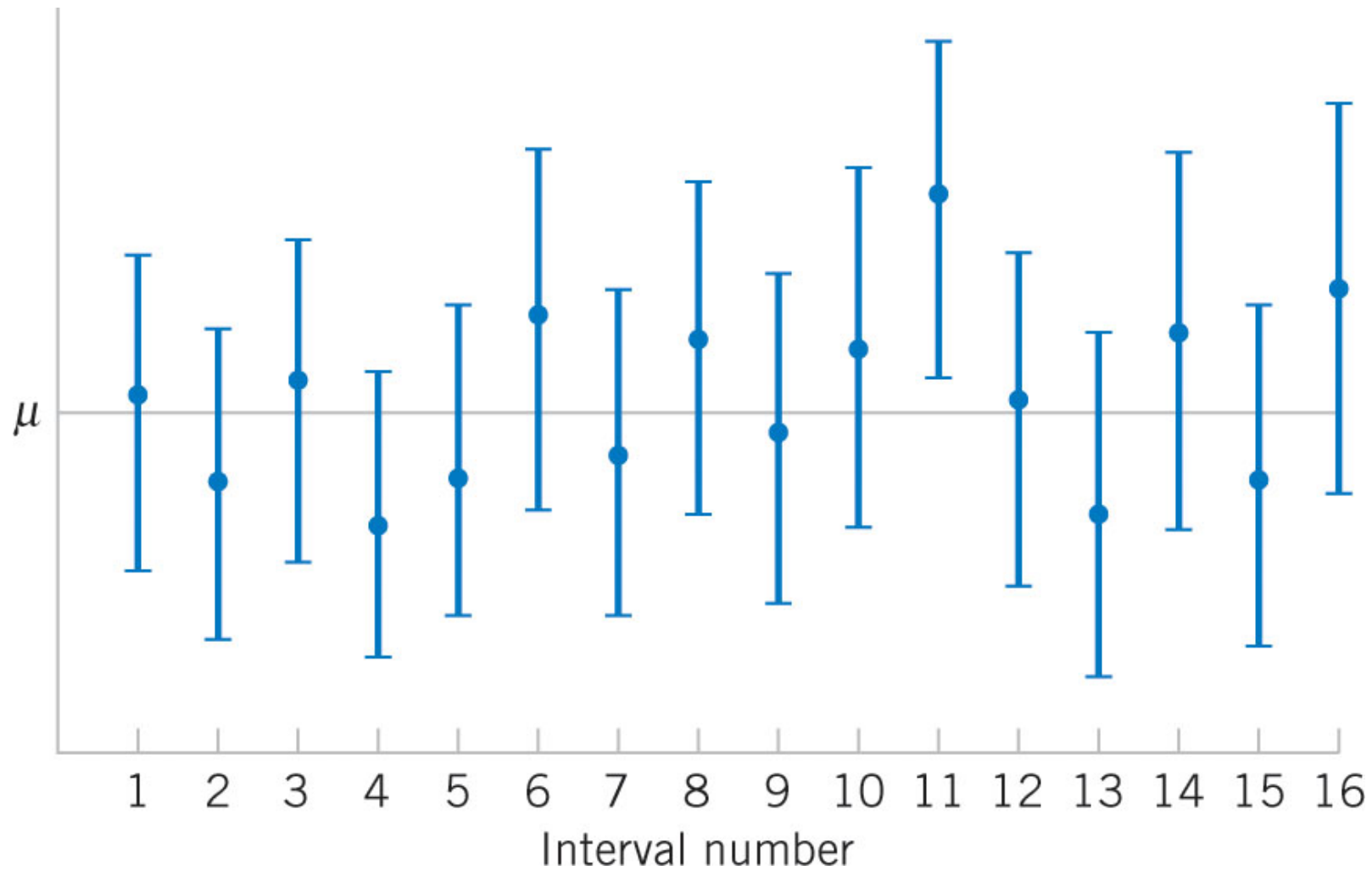
- A  $100(1-\alpha)\%$  confidence interval for a parameter  $\theta$  is a random interval  $[L, U]$ , such that
$$P(L \leq \theta \leq U) \approx 1-\alpha.$$
- $L$ : Lower confidence limit
- $U$ : Upper confidence limit



# Interpretation of confidence intervals

- If repeated random samples are taken, and a  $100(1-\alpha)\%$  confidence interval is computed for each sample, then  $100(1-\alpha)\%$  of these intervals will contain the true value of  $\theta$ .
- NOT that the CI contains the true parameter with 95% confidence!
  - Because we used a sample to calculate the point estimate and the interval

# Several CIs on a mean



## CI on Mean: $\sigma^2$ Unknown

- When  $\sigma^2$  is unknown, it can be estimated using  $s^2$
- For large sample sizes,  $n \geq 30$ , the test procedures using the test statistic can be used (CLT will ensure normal distribution of means).
- For small sample sizes,  $n < 30$ , use a  $t$ -test statistic, and further must assume that underlying distribution is normal.
- Test statistic with  $n-1$  degrees of freedom:
  - A  $100(1-\alpha)\%$  is given as:

# t-distribution

(AKA Student's t-distribution)

- The  $t$ -distribution is similar to the standard normal distribution.
- Both have mean equal to zero.
- Both are bell-shaped.
- The density  $t$ , is spread out more than the standard normal distribution.
- As  $n \rightarrow \infty$  the  $t$ -distribution approaches the standard normal distribution
- First published by William Gosset (from Guinness)



- Confidence interval for mean (known sigma)

$$\bar{X} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

- Confidence interval for mean with unknown sigma,  $n < 30$

$$\bar{X} - t_{\alpha/2, n-1} \frac{s}{\sqrt{n}} \leq \mu \leq \bar{X} + t_{\alpha/2, n-1} \frac{s}{\sqrt{n}}$$

- Confidence interval for mean with unknown sigma,  $n \geq 30$

$$\bar{X} - z_{\alpha/2} \frac{s}{\sqrt{n}} \leq \mu \leq \bar{X} + z_{\alpha/2} \frac{s}{\sqrt{n}}$$

# Inferences on the Mean of a Population (Variance Known)

- One-sided hypothesis:

$$H_0: \mu = \mu_0$$

$$H_1: \mu < \mu_0$$

Reject  $H_0$  if  $Z_0 < -z_\alpha$

$$H_0: \mu = \mu_0$$

$$H_1: \mu > \mu_0$$

Reject  $H_0$  if  $Z_0 > z_\alpha$

- Two-sided

$$H_0: \mu = \mu_0$$

$$H_1: \mu \neq \mu_0$$

*Reject  $H_0$  if  $Z_0 < -z_{\alpha/2}$  or  $Z_0 > z_{\alpha/2}$*

# Example 1

- A mining engineer is studying ways to increase the production of metal from a large copper deposit.
- A large amount of ore is bypassed because quality is too low for economical processing. The engineer is interested in a new process based on bacterial leaching.
- 100 loads of ore are removed. The final yield (in lbs of copper per ton of ore) is of interest. Based on current prices and processing costs, the break even level is 36 pounds of recoverable copper per ton.

## Example 1 continued

- What is our hypothesis?
  - $H_0$ :
  - $H_1$ :
- It was found that  $\sigma = 25lb/ton$        $\bar{x} = 40$
- For  $\alpha = 0.05$ ,  $z_{0.05} = 1.645$
- Calculate our test statistic:

## P-value

- The P-value is the lowest level of significance that would lead to the rejection of  $H_0$ . It is the probability that you would get a more extreme value than the one you got.
- That is
- $\text{P-value} \leq \alpha \rightarrow \text{reject } H_0 \text{ at level } \alpha$
- $\text{P-value} > \alpha \rightarrow \text{fail to reject } H_0 \text{ at level } \alpha$

## P-value

- For a normal distribution:

$$P = \begin{cases} 2[1 - \Phi(|z_0|)] & \text{two tailed test} \\ 1 - \Phi(z_0) & \text{upper tailed test} \\ \Phi(z_0) & \text{lower tailed test} \end{cases}$$

- The smaller the P-value, the more strongly  $H_0$  can be rejected.

# Type I error

- Hypothesis tests specify the test in terms of a **level of significance**  $\alpha$ .
- $\alpha = P(\text{Type I error})$   
 $= P(\text{reject } H_0 \mid H_0 \text{ is true})$

# Type II error

$$\beta = P(\text{Type II error})$$

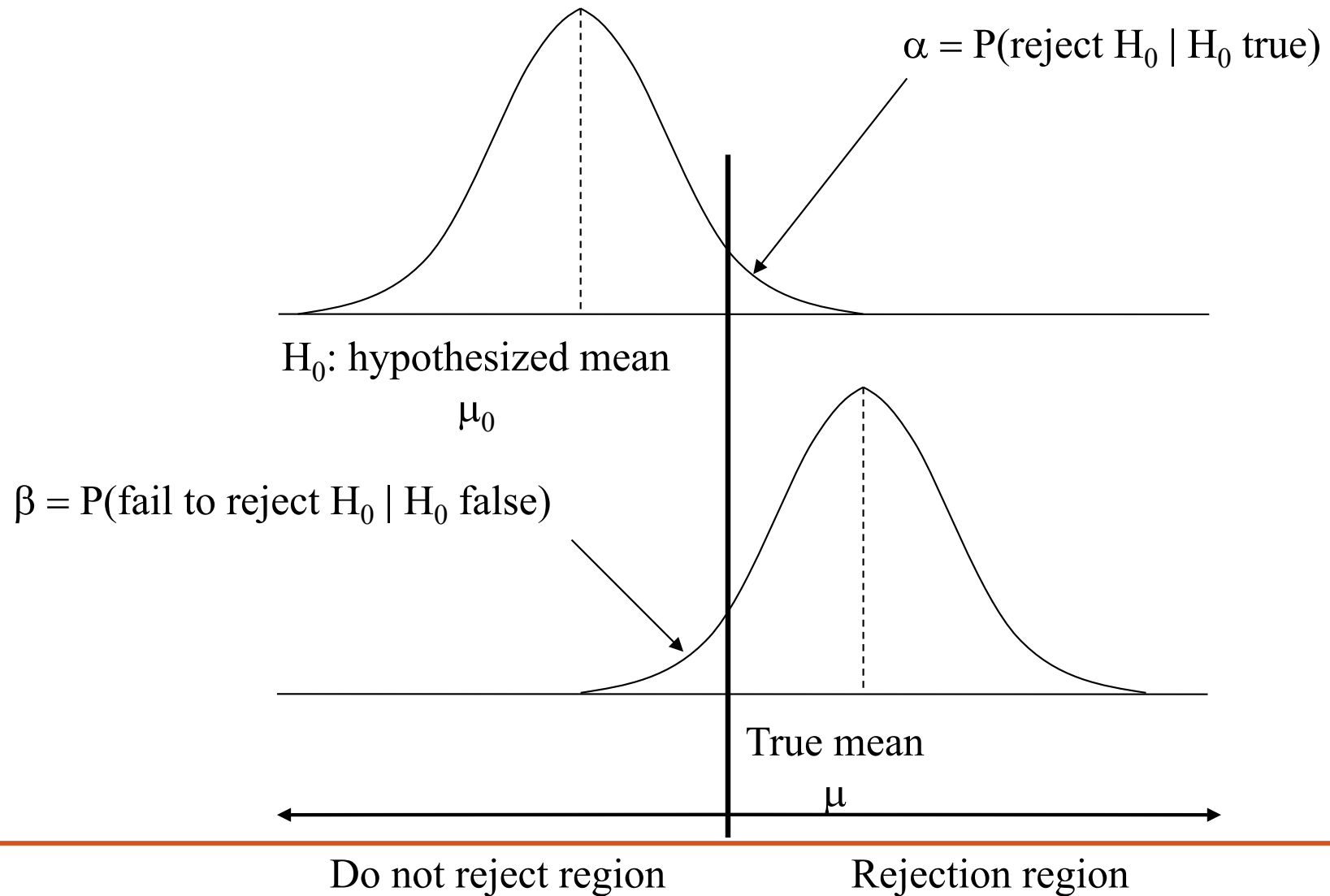
$$= P(\text{fail to reject } H_0 \text{ when } H_0 \text{ is false})$$

$$= P(\text{fail to reject } H_0 \mid H_0 \text{ is false})$$

- **Power of a test:** probability of the test leading us to reject the null hypothesis when the null hypothesis is false (equivalently, reject null hypothesis when some alternative hypothesis is true).
- Power of a test =  $1 - \beta$



# Balancing $\alpha$ and $\beta$



# Test of the Mean of a Normal Distribution with **Unknown Variance**

- Since we have **unknown variance** we will need to use the test statistic  $t_0$
- Test statistic:

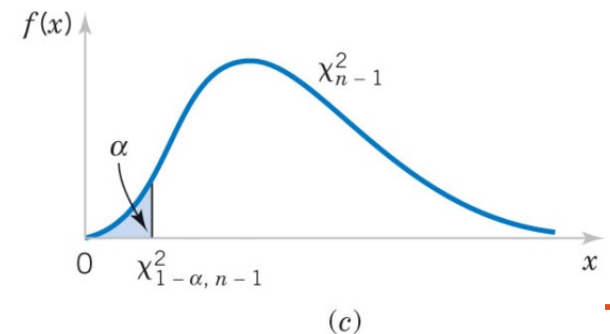
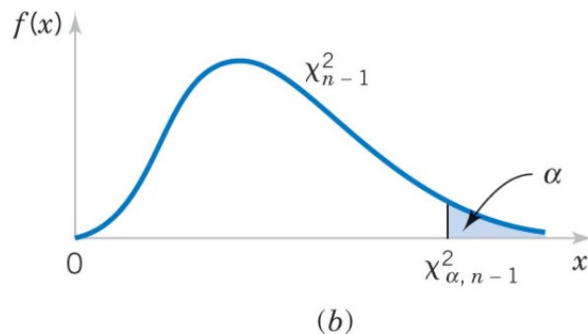
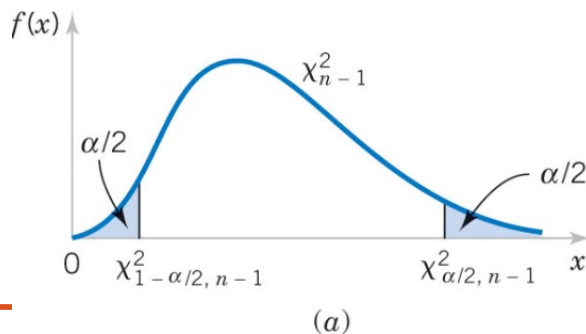
$$t_0 = \frac{\bar{X} - \mu_0}{s / \sqrt{n}}$$

## Example 2

A new steel plant has gone into production and markets a specific grade of steel. It is necessary to determine if the mean ultimate tensile strength of the new product,  $\bar{x}$  differs significantly from the industry standard  $\mu=50,000$  psi. A sample of 10 is taken from the process and yields a value of  $\bar{x}=48,500$  and standard deviation ( $s=2,700$  psi). With a significance level of  $\alpha=0.05$ , is there a significant difference between the industry standard and the new process?

# Inference on variance and standard deviation of a normal distribution

- Test the variance of a normal population compared to some specific value
- Must assume that the data follows a normal distribution
- Test statistic:  $\chi_0^2 = \frac{(n-1)S^2}{\sigma_0^2}$
- Use  $\chi^2$  distribution



# Inference on variance and standard deviation of a normal distribution

- One sided hypothesis:
  - $H_0: \sigma^2 = \sigma^2_0$                        $H_0: \sigma^2 = \sigma^2_0$
  - $H_1: \sigma^2 < \sigma^2_0$                        $H_1: \sigma^2 > \sigma^2_0$
  - Reject  $H_0$  if  $\chi^2_0 < -\chi^2_{\alpha, v-1}$                        $\chi^2_0 > \chi^2_{\alpha, v-1}$
- Two sided hypothesis:
  - $H_0: \sigma^2 = \sigma^2_0$
  - $H_1: \sigma^2 \neq \sigma^2_0$
  - Reject  $H_0$  if  $\chi^2_0 < -\chi^2_{\alpha, v-1}$  or  $\chi^2_0 > \chi^2_{\alpha, v-1}$

## Example 3

- Data from an Izod impact test resulted in a sample standard deviation of 0.25 with a sample size of 20. Does the sample standard deviation significantly differ from 0.10? Use  $\alpha=0.01$

In order to use the  $\chi^2$  statistic in hypothesis testing and confidence interval construction, we need to assume that the underlying distribution is normal.

1) The parameter of interest is the true standard deviation of Izod impact strength,  $\sigma$ . However, the answer can be found by performing a hypothesis test on  $\sigma^2$ .

2)  $H_0 : \sigma^2 = (0.10)^2$

3)  $H_1 : \sigma^2 \neq (0.10)^2$

4)  $\chi_0^2 = \frac{(n-1)s^2}{\sigma^2}$

5) Reject  $H_0$  if  $\chi_0^2 < \chi_{1-\alpha/2, n-1}^2$  where  $\alpha = 0.01$  and  $\chi_{0.995, 19}^2 = 6.8427$  or  $\chi_0^2 > \chi_{\alpha/2, n-1}^2$  where  $\alpha = 0.01$  and  $\chi_{0.005, 19}^2 = 38.58$  for  $n = 20$

6)  $n = 20, s = 0.25$

$$\chi_0^2 = \frac{(n-1)s^2}{\sigma^2} = \frac{19(0.25)^2}{(0.10)^2} = 118.75$$

7) Because  $118.75 > 38.58$  reject  $H_0$ . There is sufficient evidence to indicate the true standard deviation of Izod impact strength is significantly different from 0.10 at  $\alpha = 0.01$ .

# Inference on Population Proportion

- In a sample of  $n$ , we have  $X$  successes, we can estimate the probability of success
- Recall, if  $X$  has a binomial distribution with parameters  $n$  and  $p$ , the  $X$  can be approximated with a normal distribution that has parameters  $\mu=np_0$  and  $\sigma^2=np_0(1-p_0)$
- Test statistic:

$$Z_0 = \frac{X - np_0}{\sqrt{np_0(1 - p_0)}}$$



# Inference on Population Proportion

- One sided hypothesis:

- $H_0: p = p_0$

$$H_0: p = p_0$$

- $H_1: p < p_0$

$$H_1: p > p_0$$

- Reject  $H_0$  if  $z_0 < -z_\alpha$

$$z_0 > z_\alpha$$

- Two sided hypothesis:

- $H_0: p = p_0$

- $H_1: p \neq p_0$

- Reject  $H_0$  if  $z_0 < -z_{\alpha/2}$  or  $z_0 > z_{\alpha/2}$

# Nonparametric Tests and Approaches

# Advantages of nonparametric tests

- Usually quick and easy
- We can test categorical data, rank order data
- We can still use a specific level of significance  $100\%(1-\alpha)$
- Can be used with smaller samples than is typically required in parametric tests but are less efficient
- The most efficient test should always be used if possible.

# The sign test

- Used to test hypotheses about the median
- Key ideas:
  - In a normal distribution, 50% of the distribution is above the median, and 50% is below.
  - In a normal distribution the mean=median.
  - Test  $X_i - \text{median}_0$ , count the number of positive values ( $r^+$ )
  - $r^+$  is a binomial random variable
  - $P = P(R^+ \leq r^+, \text{ when } p=1/2)$

## Example 5

- The impurity level (in ppm) is routinely measured in an intermediate chemical process. Data:
- 2.4, 2.5, 1.7, 1.6, 1.9, 2.0, 2.5, 2.6, 2.3, 2.0, 1.8, 1.3, 1.7, 2.0, 1.9, 2.3, 1.9, 2.4, 1.6.
- Can we claim that the median impurity is less than 2.5 ppm?
- Use the sign test with  $\alpha=0.05$ , calculate the P-value for this test.

# Wilcoxon Signed-rank test

- Like the sign test, but also takes into account the magnitude of the differences

Table IX Critical Values for the Wilcoxon Signed-Rank Test

$n^*$	$\alpha$	$w_{\alpha}^*$				Two-sided tests One-sided tests
		0.10 0.05	0.05 0.025	0.02 0.01	0.01 0.005	
4						
5		0				
6		2	0			
7		3	2	0		
8		5	3	1	0	
9		8	5	3	1	
10		10	8	5	3	
11		13	10	7	5	
12		17	13	9	7	
13		21	17	12	9	
14		25	21	15	12	
15		30	25	19	15	
16		35	29	23	19	
17		41	34	27	23	
18		47	40	32	27	
19		53	46	37	32	
20		60	52	43	37	
21		67	58	49	42	
22		75	65	55	48	
23		83	73	62	54	
24		91	81	69	61	
25		100	89	76	68	

\* If  $n > 25$ ,  $W^-$  (or  $W^+$ ) is approximately normally distributed with mean  $n(n+1)/4$  and variance  $n(n+1)(2n+1)/24$ .

# Example of Wilcoxon Signed Rank Test

- We are assessing the summer salary of a company's co-op and intern students. We want to evaluate the claim that the salary is \$2000 for the summer.
- We have data from 8 co-op and interns
- 2500, 2200, 1900, 2450, 2100, 1700, 1600, 1400
- Use a Wilcoxon Signed Rank test to evaluate this claim use  $\alpha=0.05$ .

## Wilcoxon Rank-Sum test (sometimes called the Mann-Whitney test, but not really)

- Can be used to test one or two sided tests
- Procedure
  - Arrange all data in ascending order.
  - Assign ranks
  - Define  $W_1$  as smaller sample sum of the ranks
  - $W_2 = (n_1 + n_2)(n_1 + n_2 + 1)/2 - W_1$
  - Define critical values from tables



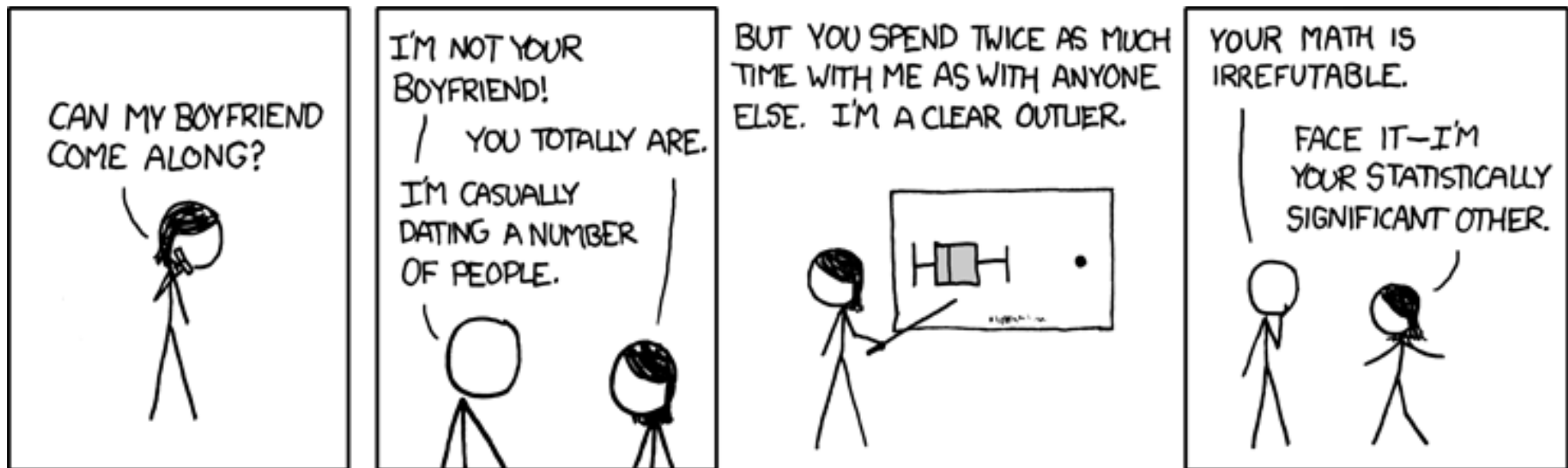
## Example 6

- A traveler travels to Seattle and uses either Delta or Alaska. Flight delays are recorded for the last six trips on each airline. Is there evidence that either airline has superior on-time arrival performance? Use  $\alpha=0.01$  and the Wilcoxon rank-sum test.
- Delta: 13, 10, 1, -4, 0, 9
- Alaska: 15, 8, 3, -1, -2, 4

 $w_{0.01}$ [illegible]

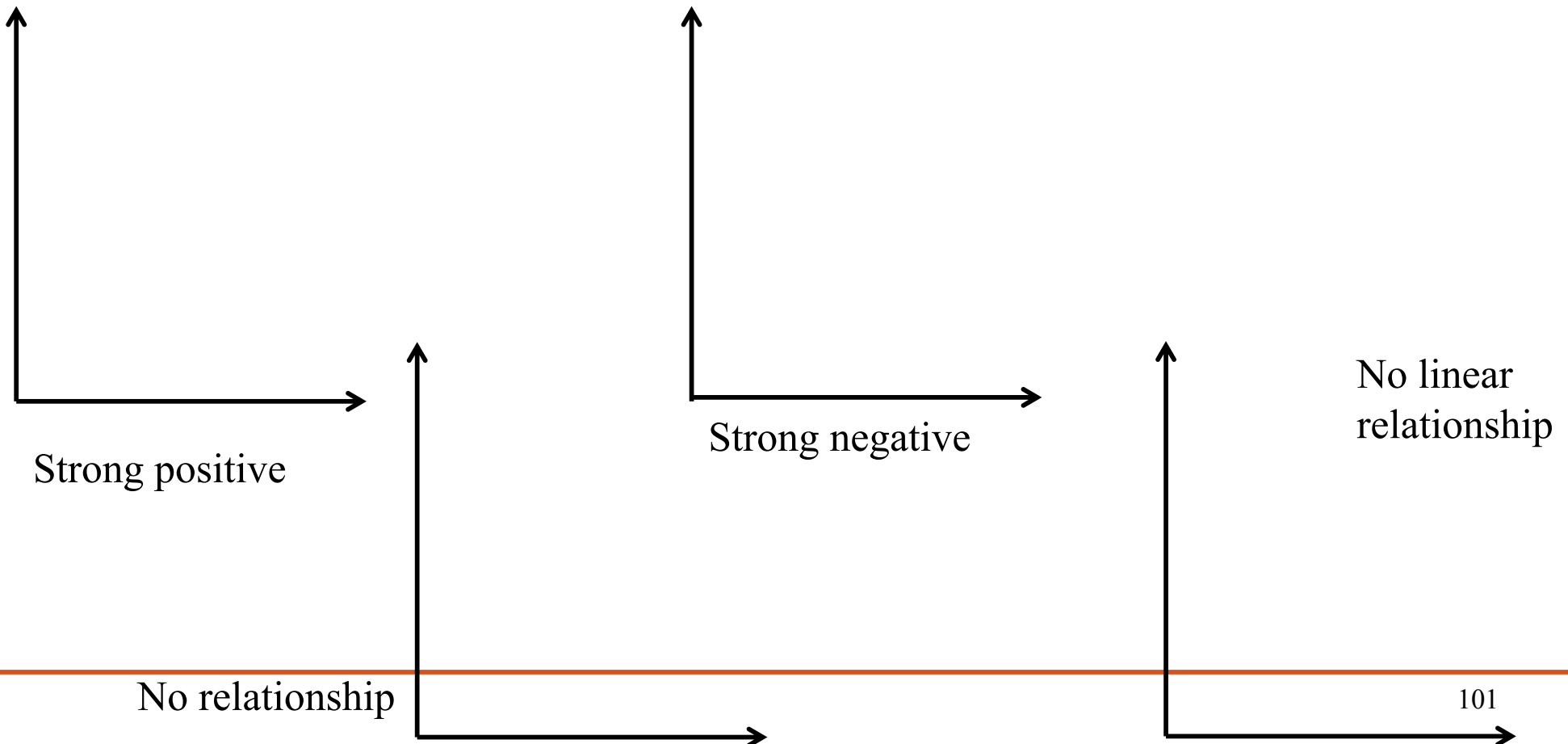
## Example 7

- Grades for exams for select students were recorded. Is there a difference in grades between Form A and Form B exams? Use  $\alpha=0.05$  and the Wilcoxon Rank-sum test.
- Form A: 24, 67, 88, 90, 98
- Form B: 25, 45, 70, 70, 80, 85,



# Correlation

- Used to show the joint behavior of two variables to determine if they are related, rather than using one to predict the other



# Sample correlation coefficient

$$r = \frac{S_{XY}}{\sqrt{\sum (x_i - \bar{x})^2} \sqrt{\sum (y_i - \bar{y})^2}} = \frac{S_{XY}}{\sqrt{S_{XX}} \sqrt{S_{YY}}}$$

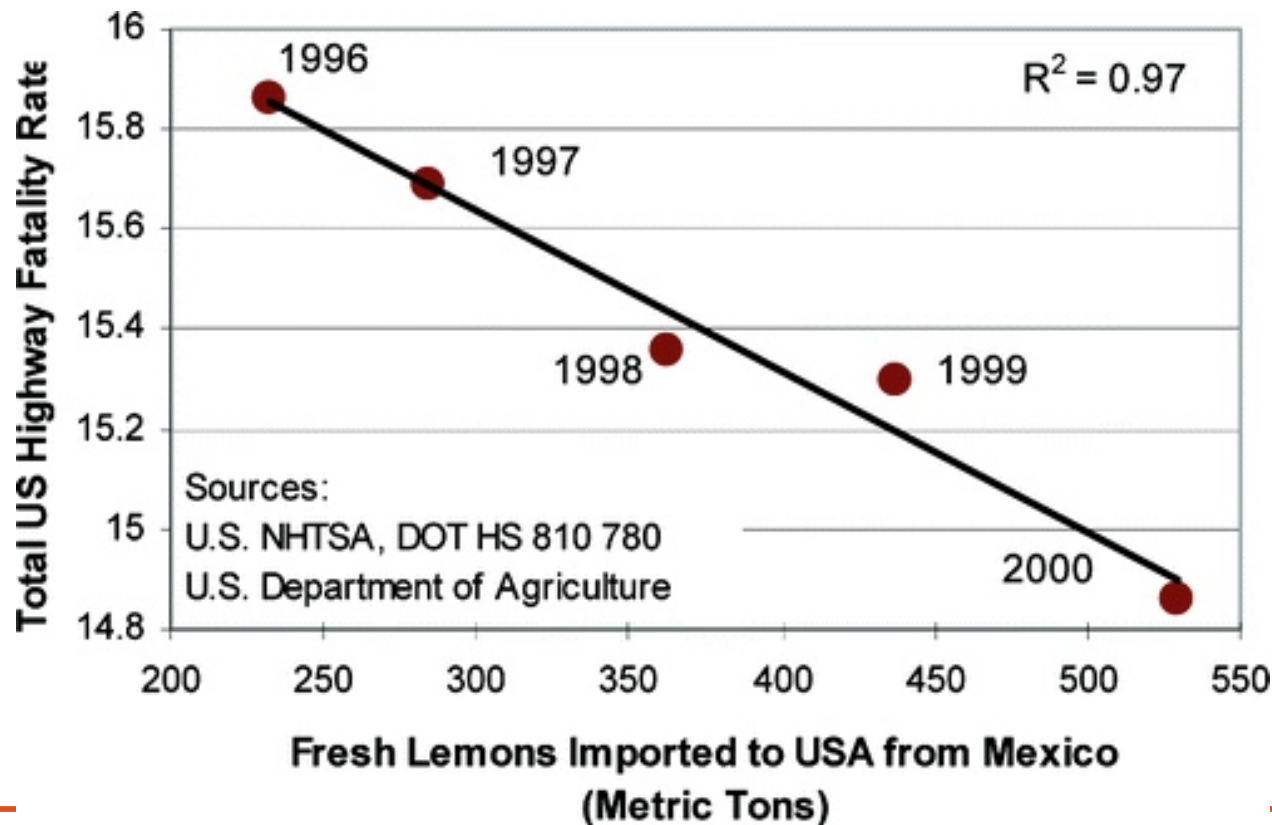
# Properties of r

- 1. The value of r does not depend on which of the two variables are labeled as x or y
- 2. The value of r is independent of the units of x and y.
- 3.  $-1 \leq r \leq 1$
- 4.  $r=1$  iff all  $(x_i, y_i)$  lie on a straight line with a positive slope, and  $r=-1$  iff all  $(x_i, y_i)$  lie on a straight line with a negative slope
- 5. The square of the sample correlation coefficient gives the value of the coefficient of determination from linear regression (next lecture)  $(r)^2=r^2$
- 6. Rule of thumb:  
 $0 \leq |r| \leq 0.5$ : weak correlation  
 $0.8 \leq |r| \leq 1.0$ : strong correlation

# Correlation vs. Causation

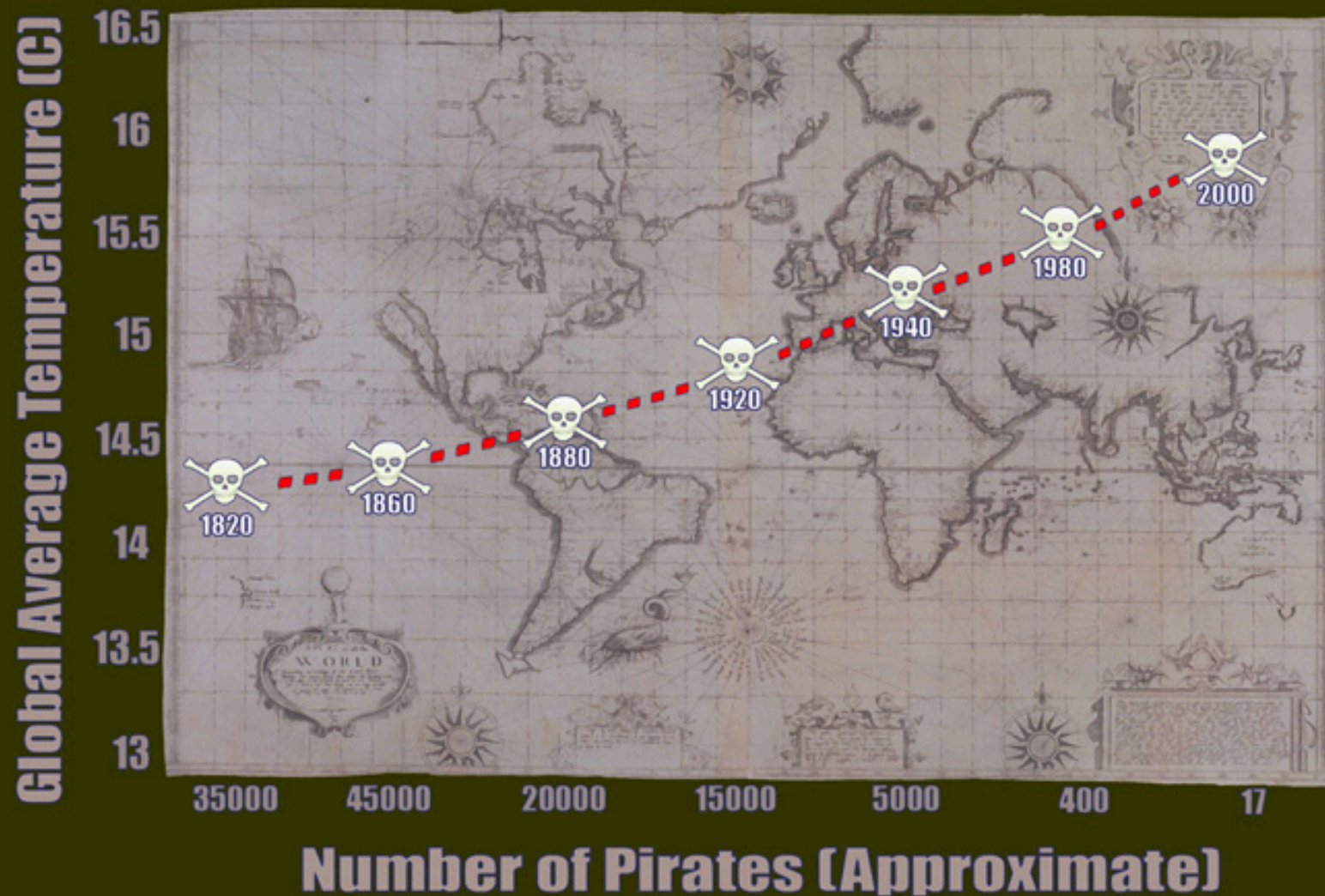
Correlation does NOT imply causation.

Empirically observed correlation is a necessary but not sufficient condition for causality. (Recall validity from earlier).

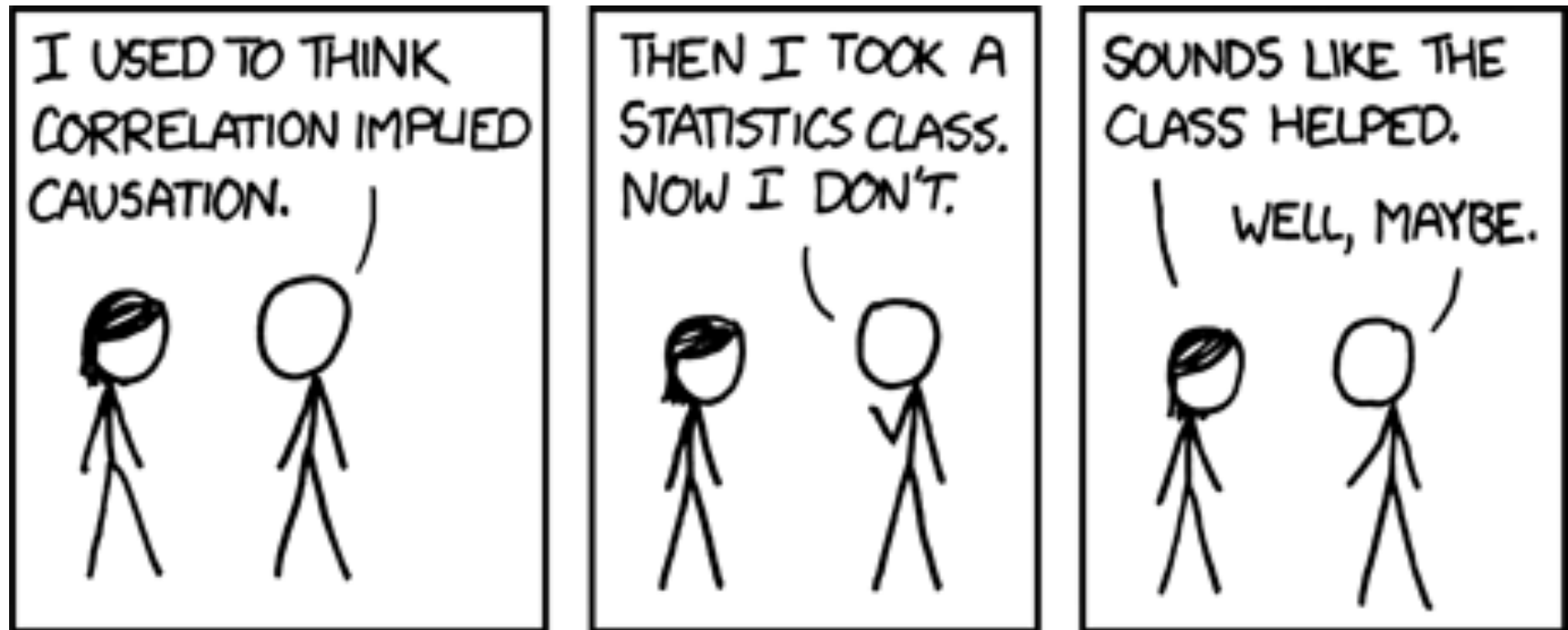




# Global Temperature Vs. Number of Pirates



<http://bama.ua.edu/~sprentic/101%20Psych%20&%20Life--Correlation-causation.htm>



# Simple linear regression

Is a statistical tool which examines the effects that some variables exert (or appear to exert) on others. It utilizes the relation (or apparent relation) between 2 or more **quantitative** variables so that 1 variable can be predicted from the others.

Examples:

Predict price of a house based on size

Predict weight based on height

Predict car mpg based on car weight

- Per Capita Health Spending and Per Capita Gross Domestic Product (GDP) in 24 OECD Countries (1989)
  - From Schieber, Poullier and Greenwald, Health Affairs, 1991

Country	Per Cap Hlth	Per Cap GDP
1 United States	2051	18.1429
2 Canada	1483	17.2857
3 Iceland	1241	15.5714
4 Sweden	1233	13.8571
5 Switzerland	1225	13.8571
6 Norway	1149	15.5714
7 France	1105	12.2857
8 Germany	1093	13.4286
9 Luxemburg	1050	14.8571
10 Netherlands	1041	13.0000
11 Austria	982	11.8571
12 Finland	949	12.8571
13 Australia	939	12.2857
14 Japan	915	13.4286
15 Belgium	879	11.8571
16 Italy	841	12.4286
17 Denmark	792	13.5714
18 UK	758	12.4286
19 New Zeland	733	10.8571
20 Ireland	561	7.8571
21 Spain	521	8.8571
22 Portugal	386	6.5714
23 Greece	337	6.4286
24 Turkey	148	4.4286

# Approach to linear regression

---

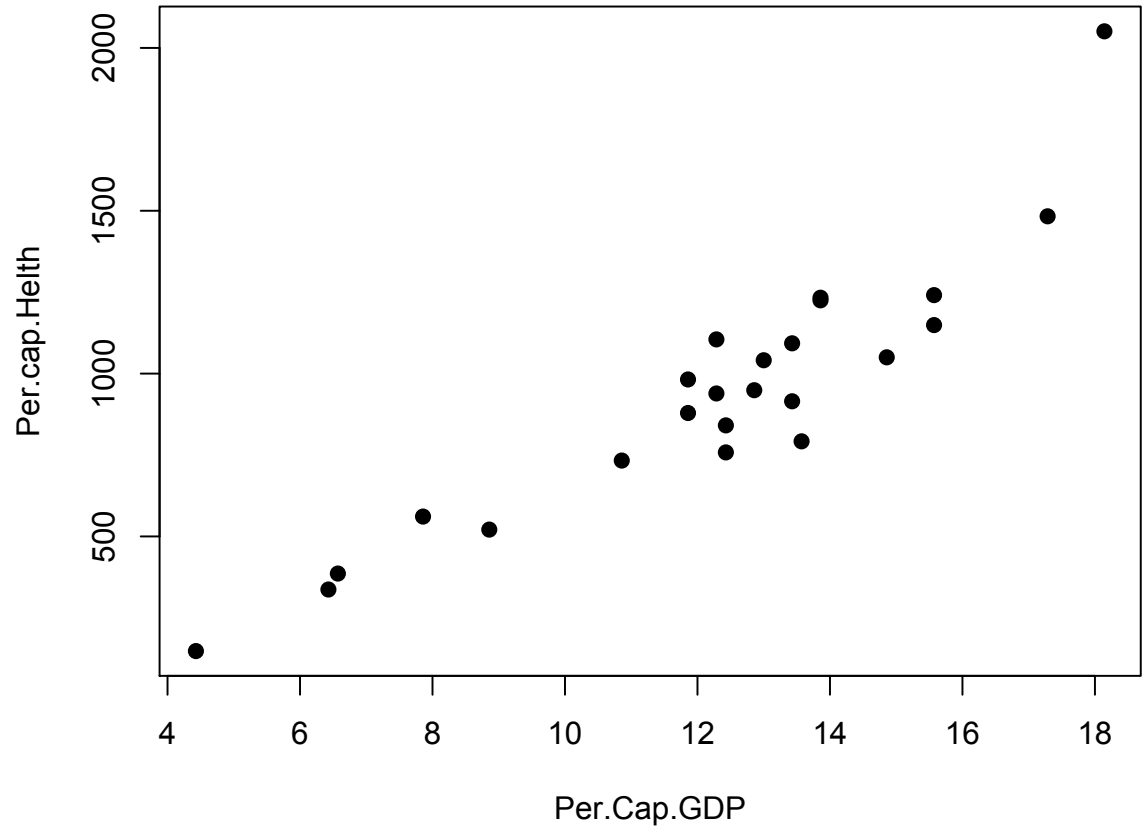
- In regression analysis, we look at the conditional distribution of the **response variable** at different levels of a **predictor variable**
- Response variable
  - also called ‘dependent’ or ‘outcome’ variable
  - what we want to explain or predict
  - in simple linear regression, response variable is continuous
- Predictor variable
  - also called ‘independent’ variable or ‘covariate’
  - in simple linear regression predictor variable is continuous
  - How we define which variable is the response and which is the predictor depends on the research question

Should always plot the data

Scatter plot

Response on Y

Predictor on X



Roughly linear

Makes sense to summarize the relationship between these variables with a straight line

# Linear function

- In considering the relationship between variables, there are two kinds:
  - Functional:
  - Statistical:

The simplest:

But we will use:

- $Y$  is the response variable that is a linear function of the predictor variable  $X$
- $\beta_0$  is the intercept; the value of  $Y$  when  $X=0$
- $\beta_1$  is the slope; how much  $Y$  changes when  $X$  increases by 1 unit

# Linear regression

- In linear regression  $\beta_0 + \beta_1 X$  represents the mean value of all of the Y's for a give value of X

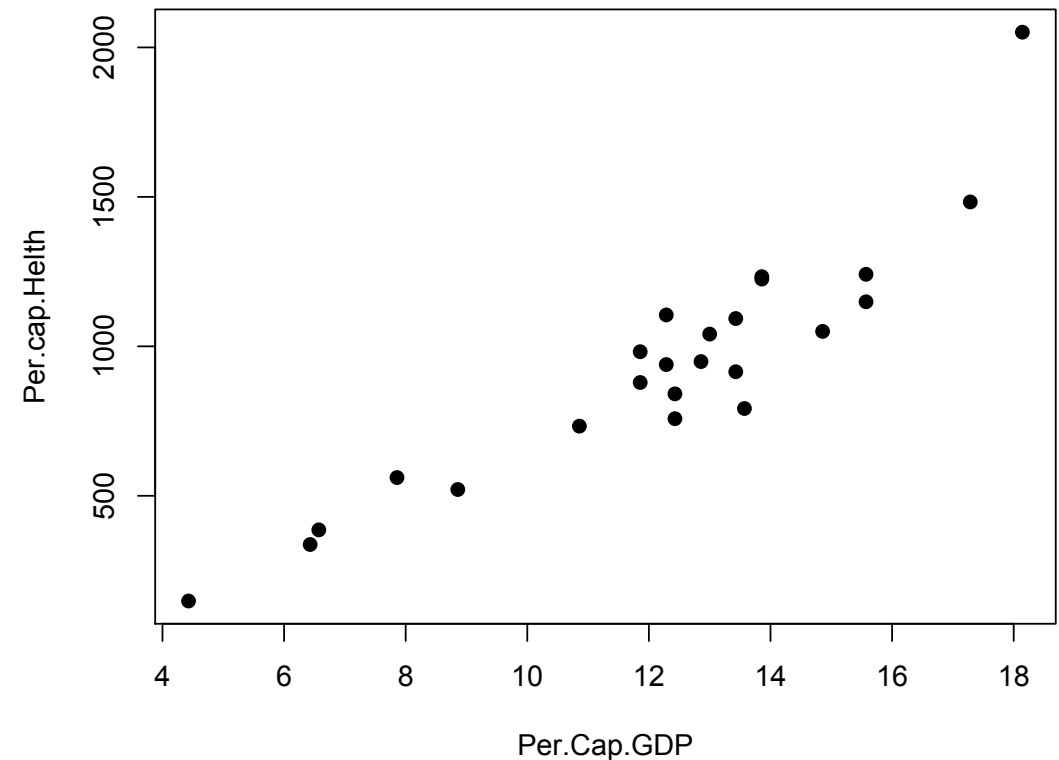
There is an entire distribution of Y values for each value of X (conditional distribution)

We say that the relationship between X and Y is linear if the means of the conditional distribution of Y|X lie on a straight line.



# Error terms

- In regression, we represent factors other than  $X_i$  that affect  $Y_i$  with an error term  $\varepsilon_i$
- Population model:

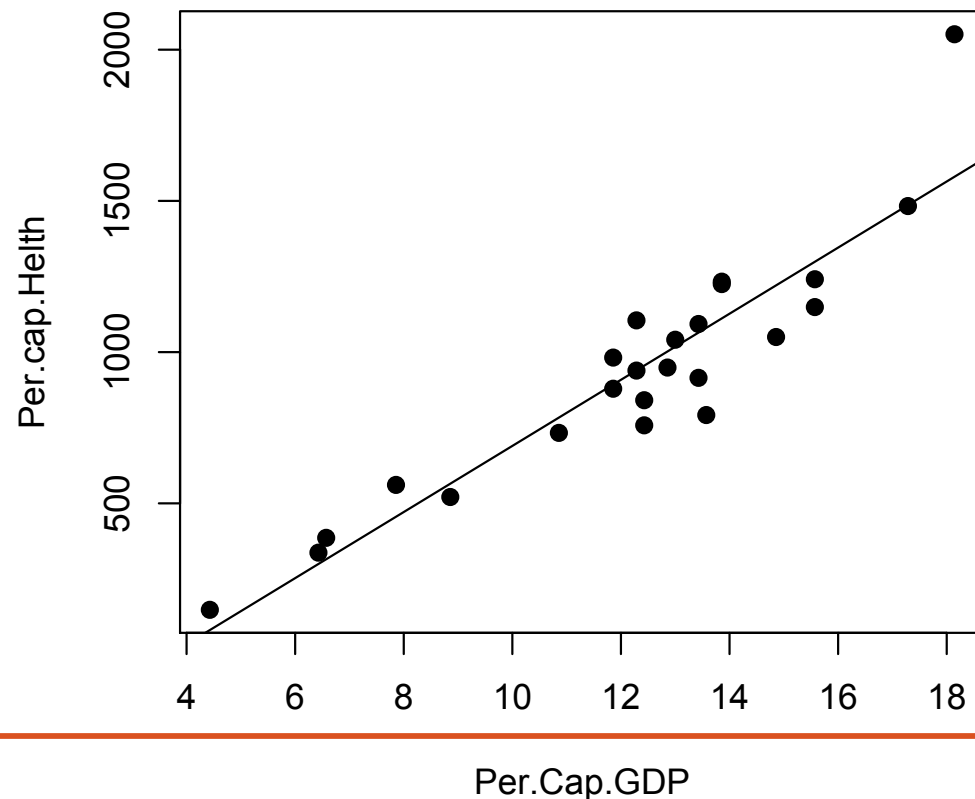


# Linear Regression

- Assumption
  - Errors  $\sim$ Normal  $(0, \sigma^2)$
- What do we know about  $y$ ?
- How do we estimate the model?

# Model

Parameter	Estimate	Std. Error	t value for H0: Parameter=0	Pr(> t )
Intercept	-402.974	121.725	-3.311	0.00318
Per.Cap.GDP	109.287	9.609	11.373	1.11e-10



# Inferences for the slope

- So far, we've been describing the relationship between two continuous variables
- Now we want to perform a hypothesis test to determine whether there is a linear relationship between two variables
  - Depends on assumptions of linear regression
- Question: Does the value of Y depend on X?
- Answer: Unless  $\beta_1 = 0$  in which case:
- Hypothesis for test for linear relationship between Y and X
- $H_0: \beta_1 = 0$   
 $H_1: \beta_1 \neq 0$

# Hypothesis testing

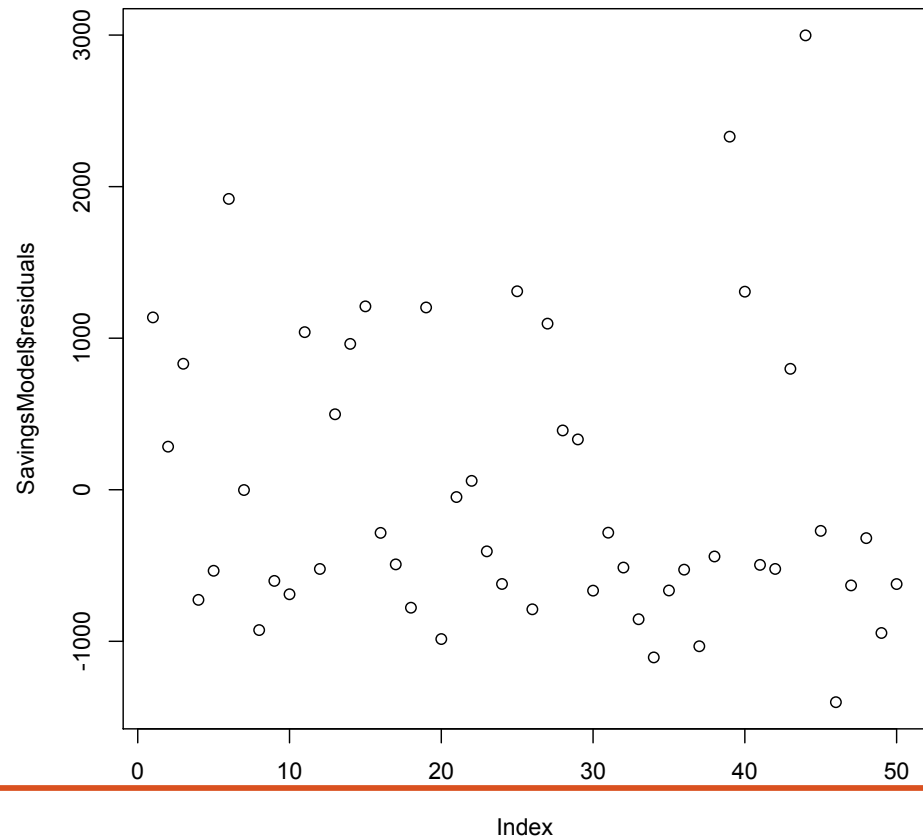
- Standard form of test statistic: estimate divided by its standard error
- Standard error of  $\hat{\beta}_1$  depends on
  - variability of Ys
  - how closely the Xs are
- Follows a t distribution with  $n-2$  degrees of freedom
- P-value: probability of obtaining a t-statistic as extreme as or more extreme than, what we got, if  $H_0$  is true.

# Hypothesis testing

- If we do not reject  $H_0$ , then we would conclude that there is no linear relationship between  $x$  and  $Y$ .
- Recall that a P-value:
  - Used in hypothesis tests to help you decide whether to reject or fail to reject a null hypothesis. The p-value is the probability of obtaining a test statistic that is at least as extreme as the actual calculated value, if the null hypothesis is true. A commonly used cut-off value for the p-value is 0.05. For example, if the calculated p-value of a test statistic is less than 0.05, you reject the null hypothesis.

# Residuals

- A linear model assumes that the residuals are normal
  - How do we test this?
  - How did we test to see if data was normally distributed before?



# Testing the residuals for normality

- Plot estimated values vs. residuals
- Plot histogram of residuals
- Plot QQ-plot
- Do a Shapiro-Wilk normality test



# From simple linear regression

- Multiple regression → more than one explanatory variable

# Designing experiments

- Statistically-based experimental design can be VERY useful for engineers

# ANOVA

Part of this lecture is based on these lecture materials:  
[http://www.utstat.toronto.edu/~olgac/sta248\\_2013/notes/sta248\\_Lecture9.pdf](http://www.utstat.toronto.edu/~olgac/sta248_2013/notes/sta248_Lecture9.pdf)

# Analysis of Variance (ANOVA)

## *Assumptions of ANOVA*

- Population distribution being sampled is normal
  - If not, may need to transform data, or use nonparametric tests
  - Can test by plotting the residuals
- The process is in control: it is repeatable
- Variance of the errors within all levels of the factor is homogeneous

*If you violate the assumptions, your results may be incorrect or misleading*

## ANOVA with single factor (one-way ANOVA)

- Completely randomized single factor experiment (completely randomized design)
  - Only one factor is varied
  - No restrictions on randomization
  - Order of experimentation is completely random

# ANOVA- single factor

Slightly different notation

Let:

$\tau_j$ : used to indicate the effect of the  $j^{\text{th}}$  level of the single factor (one treatment effect)

The model is written as:

$$Y_{ij} = \mu + \tau_j + \varepsilon_{ij}$$

$\varepsilon_{ij}$ : normally and independently distributed (NID), with random effect,  $\mu = 0$ , and  $\sigma^2$  is the same for all treatments or levels.

Where

$Y_{ij}$  represents the  $i^{\text{th}}$  observation ( $i = 1, 2, \dots, n$ ) on the  $j^{\text{th}}$  treatment ( $j = 1, 2, \dots, m$  levels).

$\mu$  is the common effect for the whole experiment (overall mean).

$\tau_j$  represents the effect of the  $j^{\text{th}}$  treatment.

$\varepsilon_{ij}$  is the random error present in the  $i^{\text{th}}$  observation on the  $j^{\text{th}}$  treatment.

# Example: data for one-way ANOVA

$H_o: \tau_j = 0$  for all  $j$ 's to be tested.

$H_a$ : at least two of the  $\tau_j$  are different.

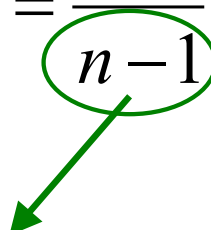
Factor Levels	Treatment (Devices <sub>j</sub> )			
	1 (auditory)	2 (visual)	3 (both)	
<b>Collected Data</b>	$Y_{11}=25$	$Y_{12}=32$	$Y_{13}=24$	
	$Y_{21}=24$	$Y_{22}=30$	$Y_{23}=23$	
	$Y_{31}=23$	$Y_{32}=24$	$Y_{33}=22$	
Totals	$T_{.1}=72$	$T_{.2}=86$	$T_{.3}=69$	$T_{..}=227$
Count	$n_1=3$	$n_2=3$	$n_3=3$	$N=9$
Means	$\bar{Y}_{.1}=24$	$\bar{Y}_{.2}=28.7$	$\bar{Y}_{.3}=23$	$\bar{Y}_{..}= 25.2$

# ANOVA- single factor calculations

Sum of Squares (Total)

$$SS_{Total} = \sum_{i=1}^n \sum_{j=1}^m (Y_{ij} - \bar{Y}_{..})^2$$

Recall the sample variance,  $s^2$

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (Y_i - \bar{Y})^2$$


For ANOVA, Degrees of freedom for TOTAL =  $N - 1$  where  $N = nm$   
and  $m$  = number of treatment groups  
 $n$  = number of samples within each treatment group



# Example: computing residuals

$H_o: \tau_j = 0$  for all  $j$ 's to be tested.

$H_a$ : at least two of the  $\tau_j$  are different.

Factor Levels	Treatment (Devices <sub>j</sub> )			
	1 (auditory)	2 (visual)	3 (both)	
cells	$Y_{11}=25$	$Y_{12}=32$	$Y_{13}=24$	
	$Y_{21}=24$	$Y_{22}=30$	$Y_{23}=23$	
	$Y_{31}=23$	$Y_{32}=24$	$Y_{33}=22$	
Totals	$T_{.1}=72$	$T_{.2}=86$	$T_{.3}=69$	$T_{..}=227$
Count	$n_1=3$	$n_2=3$	$n_3=3$	$N=9$
Means	$\bar{Y}_{.1}=24$	$\bar{Y}_{.2}=28.7$	$\bar{Y}_{.3}=23$	$\bar{Y}_{..}= 25.2$

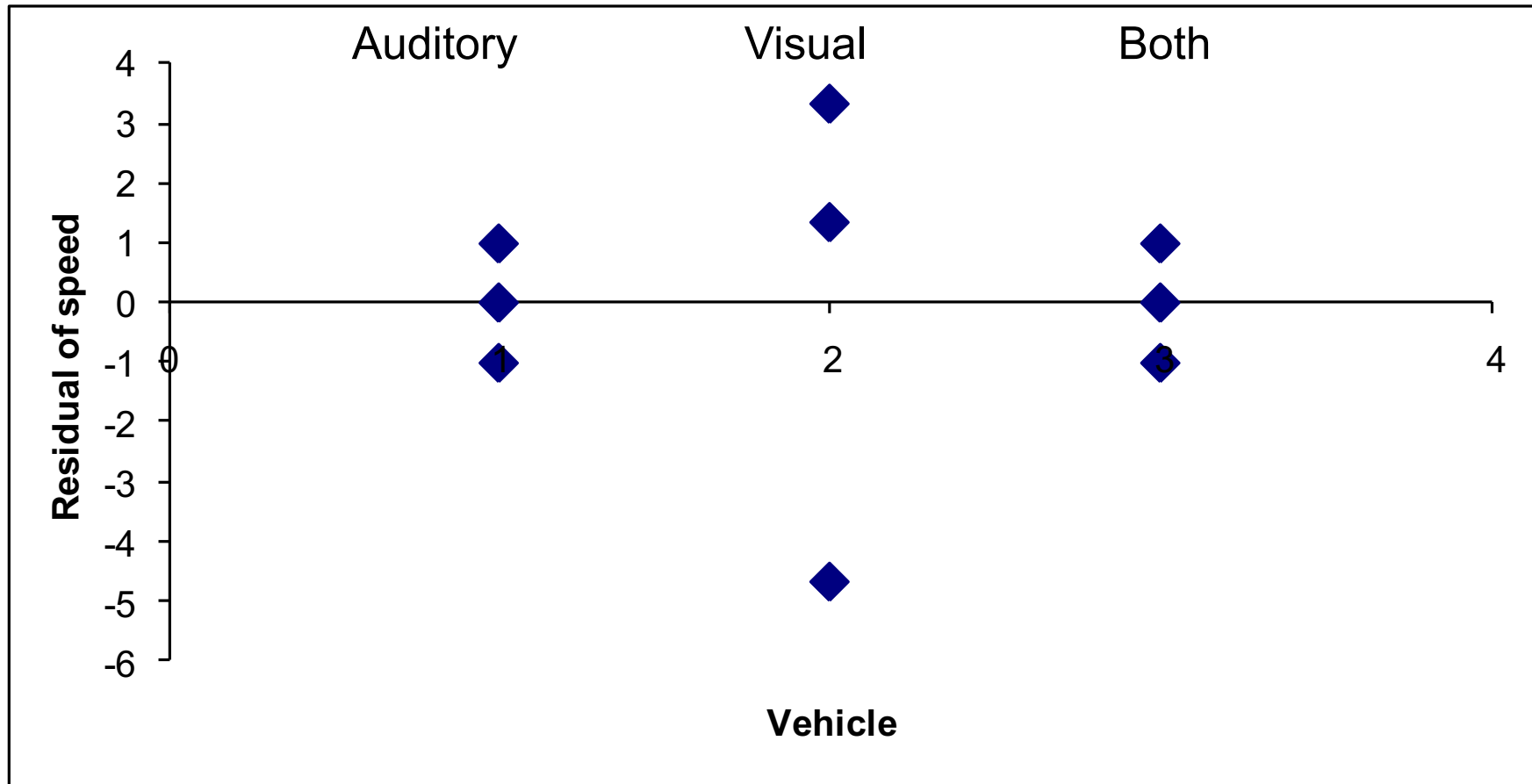
# ANOVA – single factor example

- Test for normality

$e_{ij} = y_{ij} - \bar{y}_{.j}$  For example,  $e_{11} = y_{11} - \bar{y}_{.1} = 25 - 24 = 1$

Factor Levels	Treatment (Devices <sub>j</sub> )		
	1 (auditory)	2 (visual)	3 (both)
cells	$e_{11}=1$	$e_{12}=3.33$	$e_{13}=1$
	$e_{21}=0$	$e_{22}=1.33$	$e_{23}=0$
	$e_{31}=-1$	$e_{32}=-4.67$	$e_{33}=-1$

# Test for normality



# ANOVA – single factor

	Treatment (Devices <sub>j</sub> )			
Factor Levels	1 (auditory)	2 (visual)	3 (both)	
cells	Y <sub>11</sub> =25	Y <sub>12</sub> =32	Y <sub>13</sub> =24	
	Y <sub>21</sub> =24	Y <sub>22</sub> =30	Y <sub>23</sub> =23	
	Y <sub>31</sub> =23	Y <sub>32</sub> =24	Y <sub>33</sub> =22	
Totals	T <sub>.1</sub> =72	T <sub>.2</sub> =86	T <sub>.3</sub> =69	T <sub>..</sub> =227
Count	n <sub>1</sub> =3	n <sub>2</sub> =3	n <sub>3</sub> =3	N=9
Means	$\bar{Y}_{.1}$ =24	$\bar{Y}_{.2}$ =28.7	$\bar{Y}_{.3}$ =23	$\bar{Y}_{..}$ = 25.2

Sum of Squares (Total)

$$SS_{Total} = \sum_{i=1}^n \sum_{j=1}^m Y_{ij}^2 - \frac{T_{..}^2}{N} = (25^2 + 24^2 + \dots + 23^2 + 22^2) - \frac{227^2}{9} = 93.56$$

Sum of Squares (Treatment)

$$SS_{Treatment} = \frac{1}{n} \sum_{j=1}^m T_{.j}^2 - \frac{T_{..}^2}{N} = \frac{1}{2} (72^2 + 86^2 + 69^2) - \frac{227^2}{9} = 54.89$$

# ANOVA – single factor

Source of Variation	Sum of Squares	df	Mean Square	$F_0$
A	$SS_A$	$m - 1$	$MS_A = \frac{SS_A}{m - 1}$	$F_0 = \frac{MS_A}{MS_E}$
Error (within treatment)	$SS_E$	$N - m$	$MS_E = \frac{SS_E}{N - m}$	
Total	$SS_T$	$N - 1$		

Source of Variation	Sum of Squares	df	Mean Square	$F_0$
Vehicle Display	54.89	$3 - 1 = 2$	$\frac{54.89}{2} = 27.45$	$\frac{27.45}{6.45} = 4.26$
Error (within treatment)	38.67	$9 - 3 = 6$	$\frac{38.67}{6} = 6.45$	
Total	93.56	$9 - 1 = 8$		

Do not reject

$F(\text{critical}) = F(2,6) = 5.14$

# ANOVA and regression

## ANOVA

- Developed by Ronald Fisher
- Based on univariate theory
  - To help further reduce data to more plausible levels for further experimentation

## Regression

- Developed by Karl Pearson (1908)
- Based on multivariate theory of correlation
  - Includes goodness of fit and inferences regarding coefficients

# Difference between ANOVA and regression

## ANOVA (Fitted regression model)

- *Purpose:* to seek the impact of IVs on DV
- Good: Can help us get to a better controlled measures
- Good: Can tell if differences in a particular measure (the DV) between groups (IVs) are due to chance.
- Bad: Tells us very little about the nature of that relationship
- Bad: Limited to **category** variables as IVs and a **continuous** variable as DV
- Can account for continuous DV as a covariate

## Regression for predictive model

- *Purpose:* to seek the impact of IV on DV
- Bad: Nothing is controlled
- Bad: Cannot establish causation
- Good: Determines the coefficients for each variable
- Good: Can predict the behavior of the response variable
- Good: Can use continuous IVs to predict continuous DVs

# When to use what?

- Use ANOVA when balanced dataset across categorical independent variables
- Use linear regression when:
  - datasets are not balanced (typical in surveys)
    - Note: variation explained by a factor will be different if the factor is used first (e.g., stepwise regression)
  - where the researcher can not plan the values of every explanatory variable (observational study)
  - Cannot control experimentally, so control (or adjust for) variables statistically
  - If the explanatory variables include categorical variables and continuous variables



# References

- Trochim, W.M. (2001) The Research Methods Knowledge Base. Atomic Dog Publishing. Cincinnati.
- Gonick, L. & W. Smith (1993) The Cartoon Guide to Statistics. HarperCollins, New York.