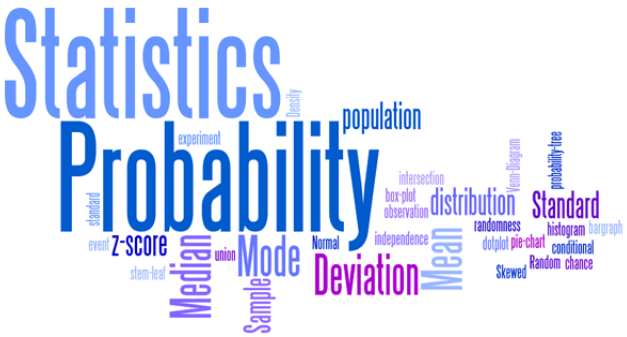
MSIT 431 Introduction to Statistics and Data Analytics

Review

Dongning Guo

Fall 2018



- Statistics is the science of data.
- Probability provides the mathematical foundation.

Outline

1.1 Data

1.2 Displaying Distributions with Graphs

1.3 Describing Distributions with Numbers

1.4 Density Curves and Normal Distributions

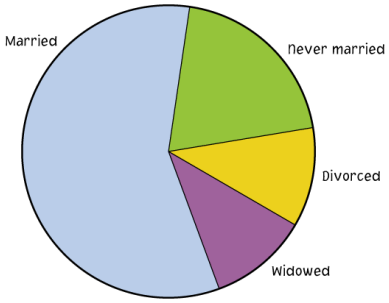
3

Terminology

- ✓ **Cases** are the objects described by a set of data. Cases may be customers, companies, experimental subjects, or other objects.
- ✓ A **variable** is a special characteristic of a case.
- ✓ A **label** is a special variable used in some data sets to distinguish between cases.
- ✓ Different cases can have different **values** of a variable.
- ❑ A **categorical** variable places each case into one of several groups, or categories.
- ❑ A **quantitative** variable takes numerical values for which arithmetic operations such as adding and averaging make sense.
- ❑ The **distribution** of a variable tells us the values that a variable takes and how often it takes each value.

4

Categorical Variables



5

Stemplots

10	0166
11	009
12	0034578
13	00359
14	08
15	00257
16	555
17	000255
18	000055567
19	245
20	3
21	025
22	0
23	
24	
25	
26	0

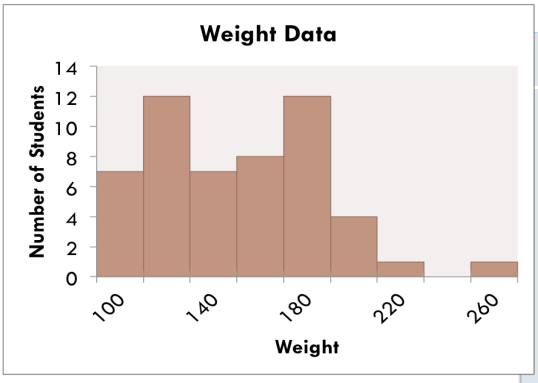
Key

20|3 means 203 pounds

Stems = 10s
Leaves = 1s

6

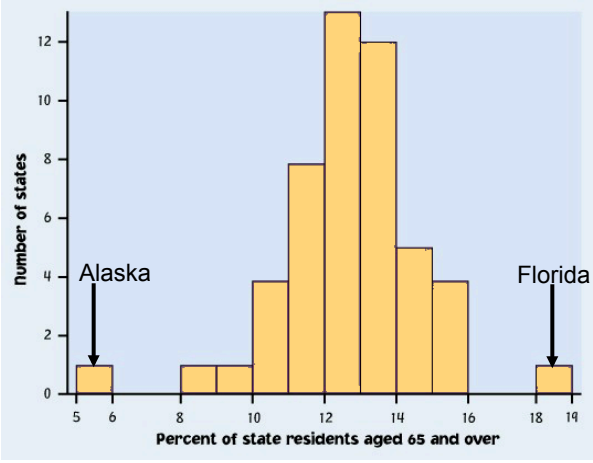
Histograms



Weight group	Count
100 - <120	7
120 - <140	12
140 - <160	7
160 - <180	8
180 - <200	12
200 - <220	4
220 - <240	1
240 - <260	0
260 - <280	1

7

Outliers



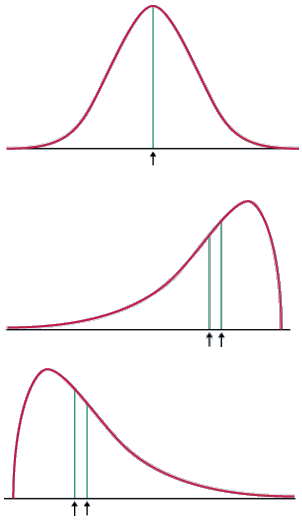
Comparing Mean and Median

The mean and median measure center in different ways, and both are useful.

The mean and median of a roughly **symmetric** distribution are close together.

If the distribution is exactly **symmetric,** the mean and median are exactly the same.

In a **skewed** distribution, the mean is usually farther out in the long tail than is the median.



9

The Five-Number Summary

The **five-number summary** of a distribution consists of the smallest observation, the first quartile, the median, the third quartile, and the largest observation, written in order from smallest to largest.

Minimum Q_1 ***M*** Q_3 ***Maximum***

10

Measuring Spread: Standard Deviation

The **standard deviation** s_x measures the average distance of the observations from their mean. It is calculated by finding an average of the squared distances and then taking the square root. This average squared distance is called the **variance.**

$$\text{variance} = s_x^2 = \frac{(x_1-\bar{x})^2+(x_2-\bar{x})^2+\ldots+(x_n-\bar{x})^2}{n-1} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2$$

$$\text{standard deviation} = s_x = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2}$$

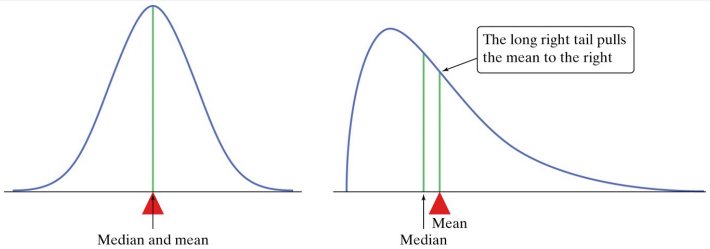
11

Density Curves

Our measures of center and spread apply to density curves as well as to actual sets of observations.

Distinguishing the Median and Mean of a Density Curve

- The **median** of a density curve is the equal-areas point—the point that divides the area under the curve in half.
- The **mean** of a density curve is the balance point, that is, the point at which the curve would balance if made of solid material.
- The median and the mean are the same for a symmetric density curve. They both lie at the center of the curve. The mean of a skewed curve is pulled away from the median in the direction of the long tail.

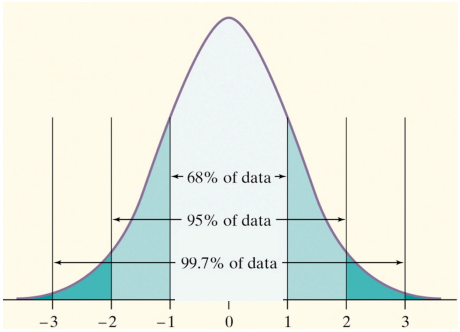


12

The 68-95-99.7 Rule

In the Normal distribution with mean μ and standard deviation σ:

- Approximately **68%** of the observations fall within σ of μ.
- Approximately **95%** of the observations fall within 2σ of μ.
- Approximately **99.7%** of the observations fall within 3σ of μ.



13

Outline

2.1 Relationships

2.2 Scatterplots

2.3 Correlation

2.4 Least-Squares Regression

2.5 Cautions about Correlation and Regression

2.6 Data Analysis for Two-Way Tables

2.7 The Question of Causation

14

Measuring linear association

The **correlation *r*** measures the strength of the linear relationship between two quantitative variables:

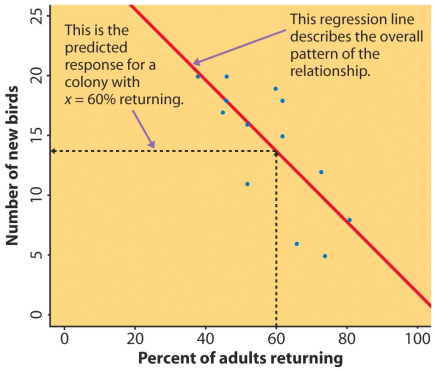
$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$$

15

Regression Line

A **regression line** is a straight line that describes how a response variable *y* changes as an explanatory variable *x* changes.

We can use a regression line to predict the value of *y* for a given value of *x*.



16

Least-Squares Regression Line

- The **least-squares regression line of *y* on *x*** is the line that minimizes the sum of the squares of the vertical distances of the data points from the line.
- If we have data on an explanatory variable *x* and a response variable *y*, the equation of the least-squares regression line is:

$$y = b_0 + b_1 x$$

where

$$b_1 = r\, s_y / s_x$$

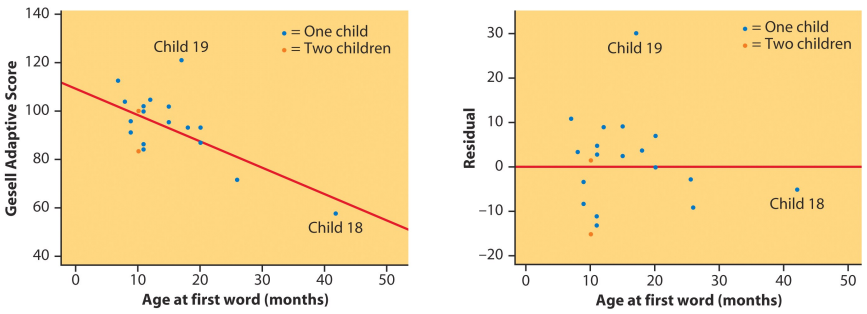
$$b_0 = \bar{y} - b_1 \bar{x}$$

17

Residual Plots

A **residual** is the difference between an observed value of the response variable and the value predicted by the regression line:

residual = observed *y* – predicted *y*



18

Two-way table and Simpson's paradox

Counts	Accepted	Not accepted	Total
Men	198	162	360
Women	88	112	200
Total	286	274	560

Percents	Accepted	Not accepted
Men	**55%**	45%
Women	44%	56%

19

Chapter 3 Producing Data

Introduction

3.1 Sources of Data

3.2 Design of Experiments

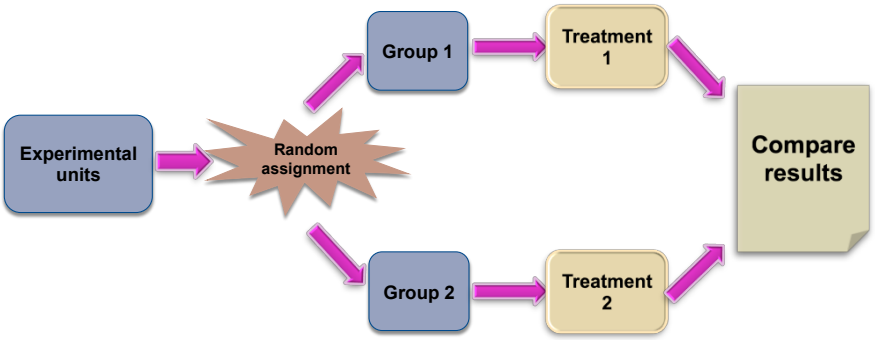
3.3 Sampling Design

3.4 Ethics

20

Randomized comparative experiments

The remedy for confounding is to perform a **comparative experiment**. In a **completely randomized design,** the treatments are assigned to all the experimental units completely by chance.

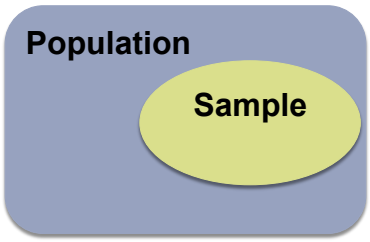


Some experiments may include a **control group** that receives an inactive treatment or an existing baseline treatment.

21

Population and Sample

- The **population** is the entire group of individuals.
- A **sample** is the part of the population from which we actually collect information.
- **Inference:** To use information from a sample to draw conclusions about the entire population.



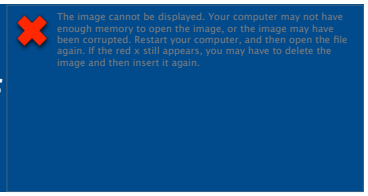
22

Simple Random Samples

A **simple random sample (SRS)** of size *n* consists of *n* individuals from the population chosen in such a way that every set of *n* individuals has an equal chance to be the sample actually selected.

23

Chapter 4 Probability: The Study of Randomness



4.1 Randomness

4.2 Probability Models

4.3 Random Variables

4.4 Means and Variances of Random Variables

4.5 General Probability Rules* (we discuss this section's material along with materials of 4.1-4.4)

24

View 1: Classical probability

- View 1: Classical probability: If an experiment has *N* equally likely outcomes, and an event *E* consists of *M* of the outcomes, then *P(E) = M/N*.
- View 2: Frequencist's view: The probability of an event is the long-term relative frequency of the event.
- View 3: Subjective probability
- View 4: Axiomatic definition of probability

25

Probability Models

Descriptions of chance behavior contain two parts: a list of possible outcomes and a probability for each outcome.

The **sample space *S*** of a chance process is the set of all possible outcomes.

An **event** is an outcome or a set of outcomes of a random phenomenon. That is, an event is a subset of the sample space.

A **probability model** is a description of some chance process that consists of two parts: a sample space *S* and a probability for each outcome.

26

Axioms of probability

Axiom 1. The probability $P(A)$ of any event A satisfies **$0 \le P(A) \le 1$.**

Any probability is a number between 0 and 1.

Axiom 2. If S is the sample space in a probability model, then **$P(S) = 1$.**

All possible outcomes together must have probability 1.

Axiom 3. If A and B are **disjoint,** **$P(A \text{ or } B) = P(A) + P(B)$.**

This is the **addition rule for disjoint events.**

If two events have no outcomes in common, the probability that one or the other occurs is the sum of their individual probabilities.

27.

Independent Events

Definition: Two events A and B are said to be **independent** if
$P(A \text{ and } B) = P(A) \times P(B)$

28

Conditional Probability

The probability that one event happens given that another event is already known to have happened is called a **conditional probability.**

When $P(A) > 0$, the probability that event B happens *given* that event A has happened is found by:

$$P(B \mid A) = \frac{P(A \text{ and } B)}{P(A)}$$

29

Bayes's Rule

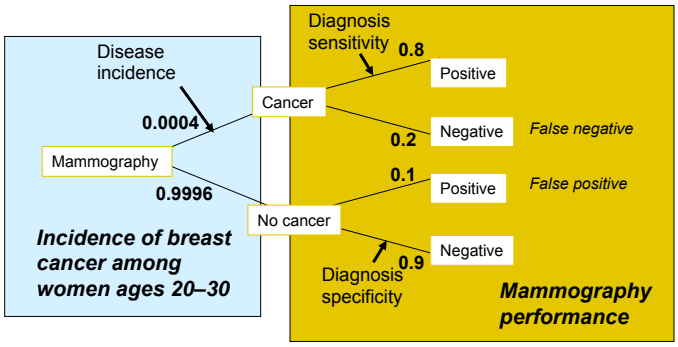
* Suppose that a sample space is decomposed into k disjoint events $A_1, A_2, \ldots, A_k$ —none of which has a 0 probability—such that $P(A_1) + P(A_2) + \ldots + P(A_k) = 1$,

- Let C be any other event such that $P(C)$ is not 0. Then

$$P(A_i \mid C) = \frac{P(C \mid A_i)P(A_i)}{P(C \mid A_1)P(A_1) + P(C \mid A_2)P(A_2) + \cdots + P(C \mid A_k)P(A_k)}$$

30

Example



If a woman in her 20s gets screened for breast cancer and receives a positive test result, what is the probability that she does have breast cancer?

$$P(\text{cancer}|\text{pos}) = \frac{P(\text{pos}|\text{cancer})P(\text{cancer})}{P(\text{pos} \mid \text{cancer})P(\text{cancer}) + P(\text{pos} \mid \text{no cancer})P(\text{no cancer})}$$

$$= \frac{0.8(0.0004)}{0.8(0.004) + 0.1(0.9996)} \approx 0.3\%$$

31

Discrete Random Variable

A **discrete random variable *X*** takes a fixed set of possible values with gaps between. The probability distribution of a discrete random variable *X* lists the values x_i and their probabilities p_i:

Value:	x_1	x_2	x_3	…
Probability:	p_1	p_2	p_3	…

The probabilities p_i must satisfy two requirements:

1. Every probability p_i is a number between 0 and 1.

2. The sum of the probabilities is 1.

To find the probability of any event, add the probabilities p_i of the particular values x_i that make up the event.

32

Continuous Random Variable

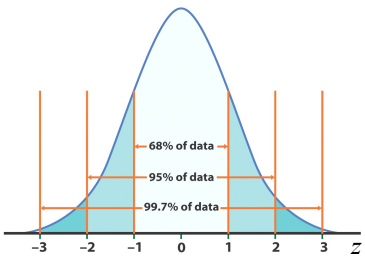
A **continuous random variable** *Y* takes on all values in an interval of numbers. The probability distribution of *Y* is described by a **density curve.** The probability of any event is the area under the density curve and above the values of *Y* that make up the event.

33

Normal Probability Models

- Often, the density curve used to assign probabilities to intervals of outcomes is the Normal curve.
- Probabilities can be assigned to intervals of outcomes using the Standard Normal probabilities in Table A.
- We **standardize** normal data by calculating *z*-scores so that any Normal curve $N(\mu,\sigma)$ can be transformed into the standard Normal curve $N(0,1)$.

$$z = \frac{x - \mu}{\sigma}$$



34

The Mean of a Random Variable

Mean of a Discrete Random Variable

Suppose that X is a discrete random variable whose probability distribution is

Value:	x_1	x_2	x_3	…
Probability:	p_1	p_2	p_3	…

To find the **mean (expected value)** of X, multiply each possible value by its probability, then add all the products:

$$E(X) = x_1p_1 + x_2p_2 + x_3p_3 + \ldots$$
$$= \sum x_ip_i$$

35

Variance of a Random Variable

Variance of a Discrete Random Variable

Suppose that X is a discrete random variable whose probability distribution is:

Value:	x_1	x_2	x_3	…
Probability:	p_1	p_2	p_3	…

and that μ_X is the mean of X. The **variance** of X is:

$$Var(X) = \sigma_X^2 = (x_1 - \mu_X)^2 p_1 + (x_2 - \mu_X)^2 p_2 + (x_3 - \mu_X)^2 p_3 + \ldots$$
$$= \sum (x_i - \mu_X)^2 p_i$$

36

The law of large numbers

- Theorem: Suppose $X_1, X_2, \ldots$ are independent and identically distributed. Suppose $\mathrm{E}(X_1)$ is finite. Then as n increase to infinity, we have the following convergence result

$$\frac{1}{n}(X_1 + X_2 + \cdots + X_n) \to \mathrm{E}(X_1)$$

37

Chapter 5
Sampling Distributions

5.1 Toward Statistical Inference

5.2 The Sampling Distribution of a Sample Mean

5.3 Sampling Distributions for Counts and Proportions

38

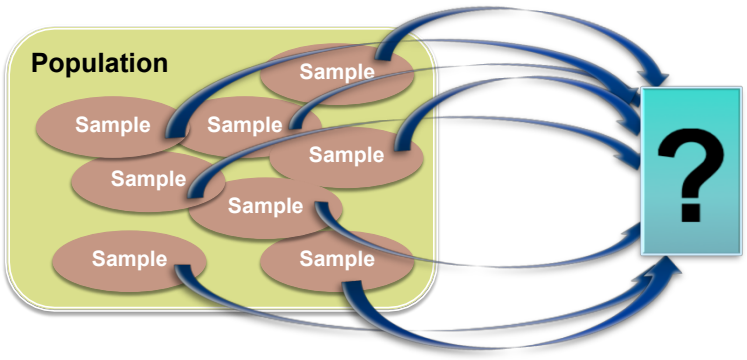
Parameters and Statistics

A **parameter** is a number that describes some characteristic of the population. In statistical practice, the value of a parameter is not known because we cannot examine the entire population.

A **statistic** is a number that describes some characteristic of a sample. The value of a statistic can be computed directly from the sample data. We often use a statistic to estimate an unknown parameter.

39

Sampling Variability



40

Sampling Distributions

The **population distribution** of a variable is the distribution of values of the variable among all individuals in the population.

The **sampling distribution** of a statistic is the distribution of values taken by the statistic in all possible samples of the same size from the same population.

41

The sampling distribution of a sample mean

The Sampling Distribution of Sample Means

Suppose that $\bar{x}$ is the mean of an SRS of size n drawn from a large population with mean μ and standard deviation σ. Then:

The **mean** of the sampling distribution of $\bar{x}$ is $\mu_{\bar{x}} = \mu$

The **standard deviation** of the sampling distribution of $\bar{x}$ is

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

42

The Central Limit Theorem

Draw an SRS of size n from any population with mean μ and finite standard deviation σ. The **central limit theorem (CLT)** says that when n is large, the sampling distribution of the sample mean $\bar{x}$ is approximately Normal:

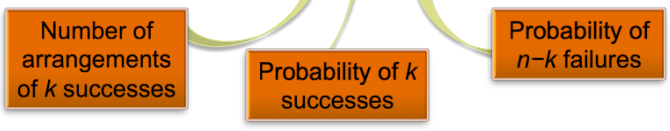
$$\bar{x} \text{ is approximately } N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

43

Binomial Probability

If X has the binomial distribution with n trials and probability p of success on each trial, the possible values of X are 0, 1, 2, …, n. If k is any one of these values,

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$



44

Sampling Distribution of a Sample Proportion

$$\hat{p} = \frac{\text{count of successes in sample}}{\text{size of sample}} = \frac{X}{n}$$

Choose an SRS of size n from a population of size N with proportion p of successes. Let $\hat{p}$ be the sample proportion of successes. Then:

The **mean** of the sampling distribution is p.

The **standard deviation** of the sampling distribution is

$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$$

For large n, $\hat{p}$ has approximately the $N(p, \sqrt{p(1-p)/n})$ distribution.

As *n* increases, the sampling distribution becomes **approximately Normal**.

45

Chapter 6 Introduction to Inference

6.1 Estimating with Confidence

6.2 Tests of Significance

6.3 Use and Abuse of Tests

6.4 Power and Inference as a Decision

46

Confidence interval

A **level C confidence interval** for a parameter has two parts:

- An **interval** calculated from the data, which has the form:

  estimate ± margin of error

- A **confidence level *C*,** where *C* is the probability that the interval will capture the true parameter value in repeated samples.

47

Inference about mean: the normal case

Assumptions:

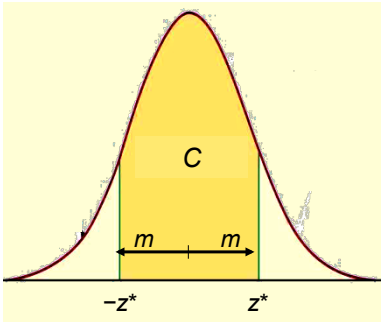
1. The variable we measure has an exactly Normal distribution $N(\mu,\sigma)$ in the population. (To be relaxed later.)
2. We don't know the population mean μ, but we do know the population standard deviation σ. (To be relaxed later.)
3. We have a simple random sample (SRS) from the population of interest.

48

The margin of error

The confidence level *C* determines the value of *z** (in Table D).

$$m = z^* \sigma / \sqrt{n}$$



49

Four steps of tests of significance

1. State the null and alternative **hypotheses.** Also pick the level of significance α.
2. Calculate the value of the **test statistic.**
3. Find the ***P*-value** for the observed data**.**
4. State a **conclusion.**

50

Test statistic

$$z = \frac{\text{estimate - hypothesized value}}{\text{standard deviation of the estimate}}$$

51

Statistical Significance

If the *P*-value is smaller than α, we say that the data are **statistically significant at level α.**

The quantity α is called the **significance level** or the **level of significance.**

P-value < α → reject H_0 → conclude H_a (in context)

P-value ≥ α → fail to reject H_0 → cannot conclude H_a (in context)

52

Chapter 7
Inference for Distributions

7.1 Inference for the Mean of a Population

7.2 Comparing Two Means

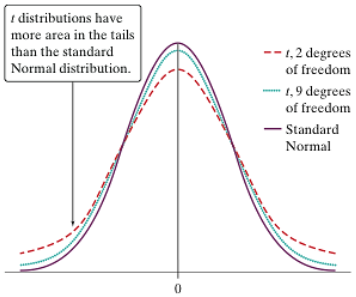
7.3 Other Topics in Comparing Distributions

The *t* distributions

Draw an SRS of size n from a large population that has a Normal distribution with mean μ and standard deviation σ. The **one-sample *t* statistic**

$$t = \frac{\bar{x} - \mu}{s_x / \sqrt{n}}$$

has the ***t* distribution** with **degrees of freedom** $df = n - 1$.



One-Sample *t* Confidence Interval

The One-Sample *t* Interval for a Population Mean

Choose an SRS of size n from a population having unknown mean μ. A level C **confidence interval** for μ is:

$$\bar{x} \pm t^* \frac{s_x}{\sqrt{n}}$$

where t^* is the critical value for the $t(n-1)$ distribution.

The **margin of error** is:

$$t^* \frac{s_x}{\sqrt{n}}$$

This interval is exact when the population distribution is Normal and approximately correct for large n in other cases.

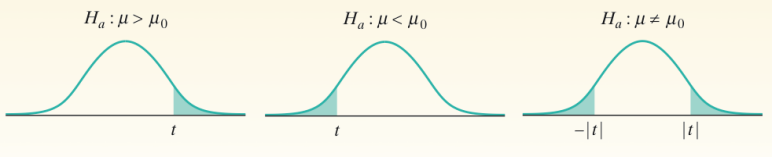
The One-Sample *t* Test

One-Sample *t* Test

Choose an SRS of size n from a large population that contains an unknown mean μ. To test the hypothesis $H_0 : \mu = \mu_0$, compute the one-sample t statistic:

$$t = \frac{\bar{x} - \mu_0}{s_x / \sqrt{n}}$$

Find the P-value by calculating the probability (at degrees of freedom = $n - 1$) of getting a t statistic this large or larger *in the direction specified by the alternative hypothesis* H_a.



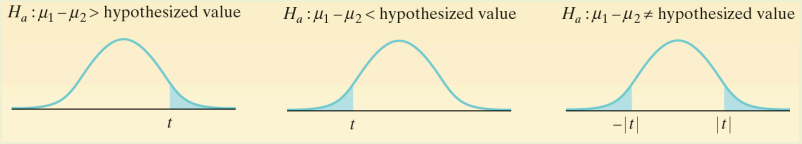
Two-Sample *t* Test

Two-Sample *t* Test for the Difference Between Two Means

Suppose the Random, Normal, and Independent conditions are met. To test the hypothesis $H_0 : \mu_1 - \mu_2 = 0$, compute the t statistic

$$t = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Find the P-value by calculating the probability of getting a t statistic this large or larger in the direction specified by the alternative hypothesis H_a. Use the t distribution with degrees of freedom approximated by technology or the smaller of $n_1 - 1$ and $n_2 - 1$.



57

Confidence Interval for μ_1 - μ_2

Two-Sample *t* Interval for a Difference Between Means

When the Random, Normal, and Independent conditions are met, a level C confidence interval for $(\mu_1 - \mu_2)$ is

$$(\bar{x}_1 - \bar{x}_2) \pm t^* \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

where t^* is the critical value at confidence level C for the t distribution with degrees of freedom either gotten from technology or equal to the smaller of $n_1 - 1$ and $n_2 - 1$.

Chapter 8
Inference for Proportions

8.1 Inference for a Single Proportion

8.2 Comparing Two Proportions

Large-Sample Confidence Interval for a Proportion

One-Sample *z* Interval for a Population Proportion

Choose an SRS of size n from a large population that contains an unknown proportion p of successes. An approximate level C **confidence interval for p** is:

$$\hat{p} \pm z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

where z^* is the critical value for the standard Normal density curve with area C between $-z^*$ and z^*.

Use this interval only when the numbers of successes and failures in the sample are both at least 15.

Plus-Four Confidence Interval

$$\tilde{p} = \frac{\text{number of successes in the sample} + 2}{n + 4}$$

Plus-Four Confidence Interval for a Proportion

$$\tilde{p} \pm z^* \sqrt{\frac{\tilde{p}(1-\tilde{p})}{n+4}}$$

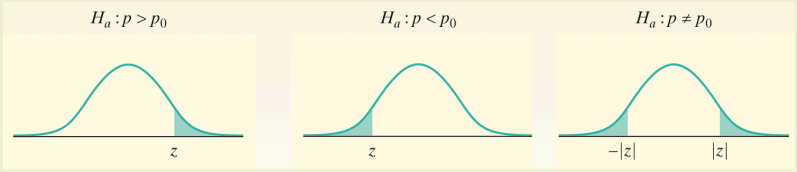
Use this interval when the confidence level is at least 90% and the sample size *n* is at least 10, with any counts of successes and failures.

Significance Test for a Proportion

Choose an SRS of size *n* from a large population that contains an unknown proportion *p* of successes. To test the hypothesis H_0: $p = p_0$, compute:

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

Find the *P*-value by calculating the probability of getting a *z* statistic this large or larger in the direction specified by the alternative hypothesis H_a:



Use this test only when the expected numbers of successes and failures are both at least 10.

Choosing the Sample Size

Sample Size for Desired Margin of Error

To determine the sample size n that will yield a level C confidence interval for a population proportion p with a maximum margin of error, solve the following:

$$n = \left(\frac{z^*}{m}\right)^2 p^*(1-p^*)$$

where p^* is a guessed value for the sample proportion. The margin of error will always be less than or equal to m if you take the guess p^* to be 0.5.

Large-Sample Confidence Interval for Comparing Proportions

Population or treatment	Parameter	Statistic	Sample size
1	p_1	$\hat{p}_1$	n_1
2	p_2	$\hat{p}_2$	n_2

Large-Sample Confidence Interval for Comparing Proportions

When the Random and Normal conditions are met, an approximate level C confidence interval for $(\hat{p}_1 - \hat{p}_2)$ is

$$(\hat{p}_1 - \hat{p}_2) \pm z^* \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

where z^* is the critical value for the standard Normal curve with area C between $-z^*$ and z^*.

Normal: The counts of "successes" and "failures" in each sample or group -- $n_1\hat{p}_1$, $n_1(1-\hat{p}_1)$, $n_2\hat{p}_2$ and $n_2(1-\hat{p}_2)$ -- are all at least 10.

Accurate Confidence Intervals for Comparing Proportions

Plus-Four Confidence Interval for Comparing Proportions

Choose independent SRSs from two large populations with proportions p_1 p_2 of successes. To get the **plus-four confidence interval for $p_1 - p_2$**, add two imaginary observations, one success and one failure, to each of the two samples. Then use the large-sample confidence interval with the new sample sizes (actual sample sizes + 2) and number of successes (actual number + 1).

$$CI: \quad (\tilde{p}_1 - \tilde{p}_2) \pm z^* \sqrt{\frac{\tilde{p}_1(1-\tilde{p}_1)}{n_1+2} + \frac{\tilde{p}_2(1-\tilde{p}_2)}{n_2+2}}$$

Use this interval when the sample size is at least 5 in each group, with any counts of successes and failures.

Significance Test for Comparing Proportions

To do a test, standardize $\hat{p}_1 - \hat{p}_2$ to get a z statistic:

$$\text{test statistic} = \frac{\text{statistic} - \text{parameter}}{\text{standard deviation of statistic}}$$

$$z = \frac{(\hat{p}_1 - \hat{p}_2) - 0}{\text{standard deviation of statistic}}$$

If H_0: $p_1 = p_2$ is true, the two parameters are the same. We call their common value p. But now we need a way to estimate p, so it makes sense to combine the data from the two samples. This **pooled** (or **combined**) **sample proportion** is:

$$\hat{p} = \frac{\text{count of successes in both samples combined}}{\text{count of individuals in both samples combined}}$$

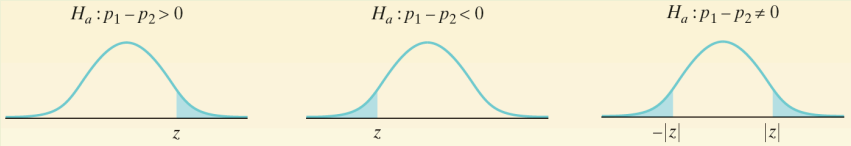
Significance Test for Comparing Proportions

Significance Test for Comparing Two Proportions

Draw an SRS of size n_1 from a large population having proportion p_1 of successes, and draw an independent SRS of size n_2 from a large population having proportion p_2 of successes. To test the hypothesis $H_0 : p_1 - p_2 = 0$, first find the pooled proportion $\hat{p}$ of successes in both samples combined. Then compute the z statistic

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

Find the P-value by calculating the probability of getting a z statistic this large or larger in the direction specified by the alternative hypothesis H_a :



Chapter 10
Inference for Regression

The image cannot be displayed. Your computer may not have enough memory to open the image, or the image may have been corrupted. Restart your computer, and then open the file again. If the red x still appears, you may have to delete the image and then insert it again.

10.1 Simple Linear Regression

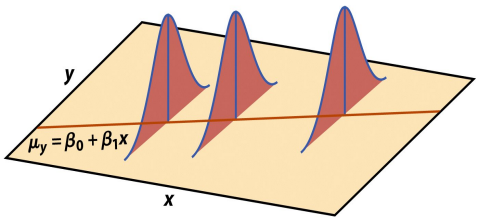
10.2 More Detail about Simple Linear Regression*

Simple Linear Regression Model

ANOVA provides information about levels of variability within a model and form a basis for tests of significance. The regression model is:

Data =	fit	+	error
y_i =	$(\beta_0 + \beta_1 x_i)$	+	(ε_i)

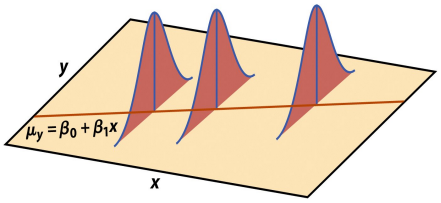
where the ε_i are **independent** and **Normally** distributed $N(0,\sigma)$, and σ is the same for all values of x.



It resembles an ANOVA, which also assumes equal variance, where

SST =	SS model	+	SS error	and
DFT =	DF model	+	DF error	

Estimating the Parameters



The **population standard deviation** σ for y at any given value of x represents the spread of the normal distribution of the ε_i around the mean μ_y.

The predicted values are $\hat{y}_i = b_0 + b_1 x_i,\ i = 1,...,n,$ and the residuals are $y_i - \hat{y}_i,\ i = 1,...,n.$

The **regression standard error, *s*,** for n sample data points is calculated from the **residuals** $(y_i - \hat{y}_i)$:

$$s = \sqrt{\frac{\sum residual^2}{n-2}} = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{n-2}}$$

s is an essentially unbiased estimate of the regression standard deviation σ.

Confidence interval for slope

The slope β_1 of the population regression line $\mu_y = \beta_0 + \beta_1 x$ is the rate of change of the mean response as the explanatory variable increases. The slope b_1 of the sample regression line is our point estimate for β_1.

The confidence interval for β_1 has the familiar form:

Estimate ± t^* · (standard deviation of estimate)

Confidence Interval for Regression Slope

A level C **confidence interval for the slope β_1** of the population regression line is:

$$b_1 \pm t^* SE_{b1}$$

Here t^* is the critical value for the t distribution with df = $n - 2$ having area C between $-t^*$ and t^*.

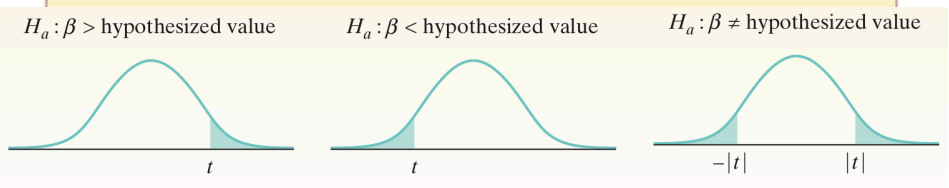
Significance test for regression slope

Significance Test for Regression Slope

To test the hypothesis H_0: β_1 = hypothesized value, compute the test statistic:

$$t = \frac{b_1 - \text{hypothesized value}}{SE_{b_1}}$$

Find the P-value by calculating the probability of getting a t statistic this large or larger in the direction specified by the alternative hypothesis H_a. Use the t distribution with df = $n - 2$.



Testing the hypothesis of no relationship

We may look for evidence of a **significant relationship** between variables *x* and *y* in the population from which our data were drawn.

For that, we can test the hypothesis that the regression slope parameter β is equal to zero.

$$H_0: \beta_1 = 0 \text{ vs. } H_a: \beta_1 \neq 0$$

Testing H_0: $\beta_1 = 0$ is equivalent to testing the **hypothesis of no correlation** between *x* and *y* in the population.

Note: *A test of hypothesis for* β_0 *is seldom of interest, mainly because* β_0 *often has no practical interpretation.*

Confidence interval for mean response

We can calculate a confidence interval for the population mean μ_y of all responses *y* when *x* takes the value *x** (within the range of data).

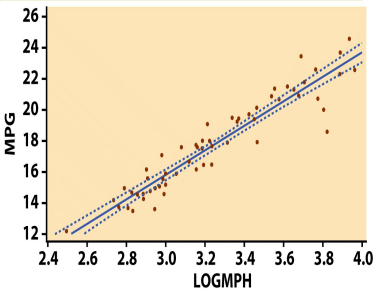
The **level *C* confidence interval for the mean response μ_y** at a given value *x** of *x* is:

$$\hat{\mu}_y \pm t^* SE_{\hat{\mu}}$$

where *t** is the value such that the area under the $t(n-2)$ density curve between $-t^*$ *and* t^* is *C.*

A separate confidence interval could be calculated for μ_y along all the values that *x* takes.

Graphically, the series of confidence intervals is shown as a continuous interval on either side of $\hat{y}$.



Prediction Intervals

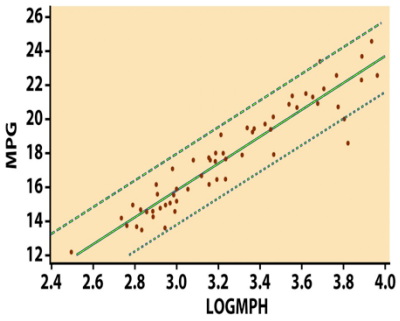
The **level *C* prediction interval for a single observation** on *y* when *x* takes the value *x** is:

$$\hat{y} \pm t^* SE_{\hat{y}}$$

*t** is the critical value for the ***t* (*n* – 2)** distribution with area C between $-t^*$ and $+t^*$.

The prediction interval accounts for error in estimating β_0 and β_1 as well as uncertainty about the value of *y* being predicted.

Graphically, the series of prediction intervals is shown as a continuous interval on either side of $\hat{y}$. These intervals are wider than the corresponding confidence intervals for μ_y.



Calculations for regression inference

To assess variation in the estimates of β_0 and β_1, we calculate the standard errors for the estimated regression coefficients.

The standard error of the slope estimate b_1 is:

$$SE_{b1} = \frac{s}{\sqrt{\sum (x_i - \bar{x})^2}}$$

The standard error of the intercept estimate b_0 is:

$$SE_{b0} = s\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum (x_i - \bar{x})^2}}$$

Calculations for regression inference

To estimate mean responses or predict future responses, we calculate the following standard errors:

The standard error of the estimate of the mean response μ_y is:

$$\mathrm{SE}_{\hat{\mu}} = s\sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum(x_i - \bar{x})^2}}$$

The standard error for predicting an individual response y is:

$$\mathrm{SE}_{\hat{y}} = s\sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum(x_i - \bar{x})^2}}$$

Inference for correlation

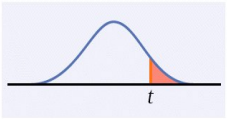
The test of significance for ρ uses the one-sample t-test for: H_0: $\rho = 0$.

We compute the t statistic for sample size n and correlation coefficient r.

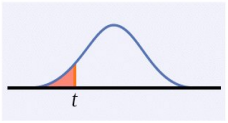
$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

The P-value is the area under $t(n-2)$ for values of T as or more extreme than t in the direction of H_a.

H_a: $\rho > 0$ is $P(T \geq t)$



H_a: $\rho < 0$ is $P(T \leq t)$



H_a: $\rho \neq 0$ is $2P(T \geq |t|)$