SISG & SISMID 2014

SISG & SISMID Module 1: Probability and Statistical Inference

19th Summer Institute in Statistical Genetics

6th Summer Institute in Statistics and Modeling in Infectious Diseases

UNIVERSITY of WASHINGTON

(This page left intentionally blank.)

Probability

Overview

- Definitions of Probability
- Sample Space, Events
- Basic Properties
- Joint, Marginal, Conditional Probability
- Rules of Probability
- Screening – Application of Bayes' Rule

NOTHING IN LIFE IS CERTAIN. IN EVERYTHING WE DO, WE GAUGE THE CHANCES OF SUCCESSFUL OUTCOMES, FROM BUSINESS TO MEDICINE TO THE WEATHER. BUT FOR MOST OF HUMAN HISTORY, *PROBABILITY*, THE FORMAL STUDY OF THE LAWS OF CHANCE, WAS USED FOR ONLY ONE THING: *GAMBLING*.

Liber de ludo aleae ("Book on Games of Chance")
by Gerolamo Cardano. Written 1526 (published 1663).
First systematic treatment of probability.
(Included section on effective cheating methods.)

Probability

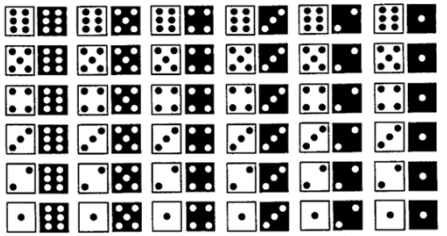
Probability provides a measure of uncertainty associated with the occurrence of events or outcomes

Definitions:

1. **Classical**: P(E) = m/N

   If an event can occur in N mutually exclusive, equally likely ways, and if m of these possess characteristic E, then the probability of E is equal to m/N.

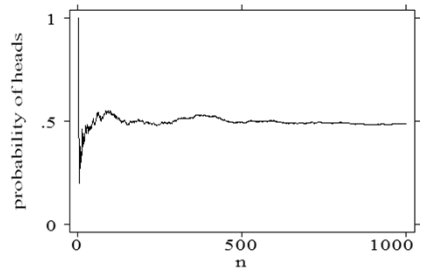
Example: What is the probability of rolling a total of 7 on two dice?



2. **Relative Frequency**: $P(E) \approx m / n$

If a process or an experiment is repeated a large number of times, n, and if the characteristic, E, occurs m times, then the relative frequency, m/n, of E will be approximately equal to the probability of E.

» Around 1900, the English statistician Karl Pearson heroically tossed a coin 24,000 times and recorded 12,012 heads, giving a proportion of 0.5005.



3. **Personal Probability**

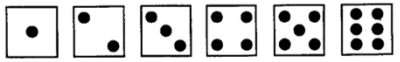
What is the probability of life on Mars?

Sample Space

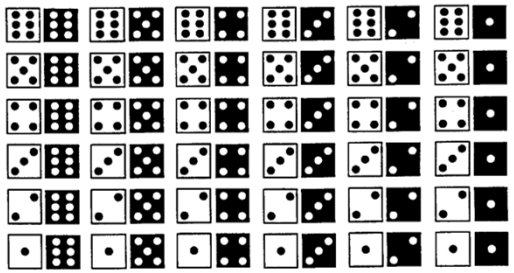
The **sample space** consists of the possible outcomes of an experiment. An **event** is an outcome or set of outcomes.

For a coin flip the sample space is (H,T).

THE SAMPLE SPACE OF THE THROW OF A *SINGLE DIE* IS A LITTLE BIGGER.



AND FOR A *PAIR* OF DICE, THE SAMPLE SPACE LOOKS LIKE THIS (WE MAKE ONE DIE WHITE AND ONE BLACK TO TELL THEM APART):



Sample Space

Key Point # 1:

Whenever you read an article with statistical results, try to identify the sample space. The sample space used by the article may not be the one they want you to think it is.

Example: Woman Wins NJ Lottery Twice

NY Times stated chance was 1 in 17 trillion. True for one particular person purchasing just one ticket each for two different runs.

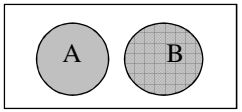
Not true if the question is: "What is the chance that ***someone*** will win the lottery twice in his/her lifetime?" Almost a sure thing!

Back in late 1980's, estimated there is about a 50% chance this will happen in 7-year period.

Diaconis, P., and F. Mosteller. (1989). Methods of Studying Coincidences. *Journal of the American Statistical Association*, **84**(408), 853-861.

Basic Properties of Probability

1. Two events, A and B, are said to be mutually exclusive (disjoint) if only one or the other, but not both, can occur in a particular experiment.



2. Given an experiment with n mutually exclusive events, E_1, E_2, …., E_n, the probability of any event is non-negative and less than 1:

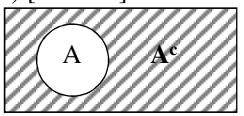
$$0 \le P(E_i) \le 1$$

3. The sum of the probabilities of an exhaustive collection (i.e. at least one must occur) of mutually exclusive outcomes is 1:

$$\sum_{i=1}^{n} P(E_i) = P(E_1) + P(E_2) + \ldots + P(E_n) = 1$$

4. The probability of all events other than an event A is denoted by $P(A^c)$ [A^c stands for "A complement"] or $P(\overline{A})$ ["A bar"]. Note that

$P(A^c) = 1 - P(A)$



Basic Properties of Probability

Example: A single die

Consider the following events:

E_1 = roll a 1

E_2 = roll an even number

E_3 = roll a 4, 5 or 6

E_4 = roll a 3 or 5

1) What is $Pr(E_4)$?

2) Are E_2 and E_3 mutually exclusive? E_2 and E_4?

3) Find a mutually exclusive, exhaustive collection of events. Do the probabilities add to 1?

4) What is $Pr(E_4^c)$?

Notation for Joint Probabilities

- If A and B are any two events then we write

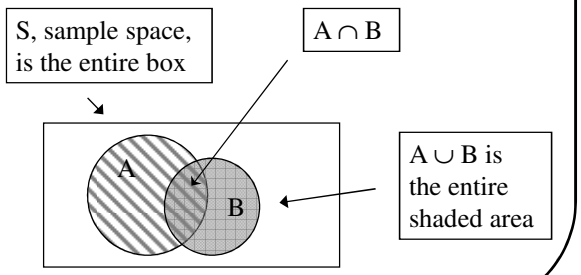
  P(A or B) *or* P(A ∪ B)

  to indicate the probability that event A <u>or</u> event B (or both) occurred.

- If A and B are any two events then we write

  P(A and B) *or* P(AB) *or* P(A ∩ B)

  to indicate the probability that both A <u>and</u> B occurred.



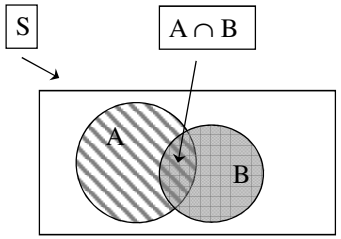
Notation for Joint Probabilities

- If A and B are any two events then we write

  P(A given B) *or* P(A|B)

  to indicate the probability of A among the subset of cases in which B is known to have occurred.

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



Conditional Probability

The conditional probability of an event A given B (i.e. given that B has occurred) is denoted P(A | B).

		Disease Status		
		Pos.	Neg.	
Test Result	Pos.	9	80	89
	Neg.	1	9910	9,911
		10	9990	10,000

What is P(test positive)?

What is P(test positive | disease positive)?

What is P(disease positive | test positive)?

Example - Joint Probabilities

2.6.2. The following table shows the first 1000 patients admitted to a clinic for retarded children by diagnostic classification and level of intelligence. For this group find:

(a) $P(A_3 \cap B_4)$.
(b) The probability that a patient picked at random is severely retarded.
(c) The probability that a patient picked at random is either not retarded or is borderline.
(d) The probability that a patient picked at random is profoundly retarded and has Down's syndrome.
(e) The probability that a patient is profoundly retarded, given that he has Down's syndrome.

	Level of Retardation						
Major Diagnostic Classification	A_1 Not Retarded	A_2 Pro-found	A_3 Severe	A_4 Moderate	A_5 Mild	A_6 Border-line	Total
B_1 Encephalopathies	33	38	57	114	103	55	400
B_2 Down's syndrome	2	4	34	88	27	5	160
B_3 Congenital cerebral defect	10	2	6	6	6	0	30
B_4 Mental retardation of unknown cause	0	0	9	36	62	35	142
B_5 Other	161	0	8	16	8	75	268
Total	206	44	114	260	206	170	1000

Joint Probabilities

Key Point # 2:

A probability depends on your definition of the sample space.

The sample space changes with knowledge of the circumstances or what has occurred.

Example:

Car insurance companies don't set rates based on using probabilities based on ALL drivers, they use probabilities based on categories of drivers (e.g., Male <22, Female 40-49, etc.)

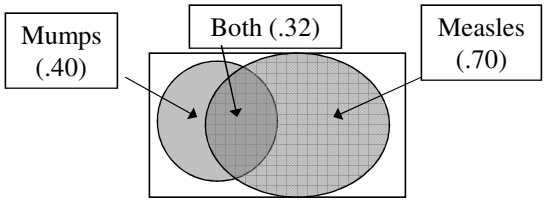
General Probability Rules

- Addition rule

If two events A and B are not mutually exclusive, then the probability that event A or event B occurs is:

$$P(A \text{ or } B) = P(A) + P(B) - P(AB)$$

E.g. Of the students at Anytown High school, 40% have had the mumps, 70% have had measles and 32% have had both. What is the probability that a randomly chosen student has had at least one of the above diseases?



General Probability Rules

- Multiplication rule (special case – independence)

If two events, A and B, are "independent" (probability of one does not depend on whether the other occurred) then

$$P(AB) = P(A)P(B)$$

E.g.

Suppose
P(mumps) = 40%

P(measles) = 70%

If independent, then we predict
P(mumps, measles) = .4*.7=.28

Easy to extend for independent events A,B,C,…

$$P(ABC\ldots) = P(A)P(B)P(C)\ldots$$

General Probability Rules

Two events A and B are said to be <u>independent</u> if and only if

$$P(A|B) = P(A) \text{ or}$$
$$P(B|A) = P(B) \text{ or}$$
$$P(AB) = P(A)P(B).$$

(Note: If any one holds then all three hold)

<u>E.g.</u>
Suppose
P(mumps) = .4, P(measles) = .7
P(both) = .32.
Are the two events independent?

No, because P(mumps and measles) = .32 while
P(mumps) P(measles) = .28

The notion of independent events is pervasive throughout statistics …

General Probability Rules

- <u>Multiplication rule (general)</u>

  More generally, however, A and B may not be independent. The probability that one event occurs may depend on the other event. This brings us back to conditional probability. The general formula for the probability that both A and B will occur is

$$P(AB) = P(A \mid B)P(B) = P(B \mid A)P(A)$$

<u>E.g.</u>

Suppose
P(mumps) = 40%
P(measles | mumps) = 80%

then

P(both) = .80*.40 = .32

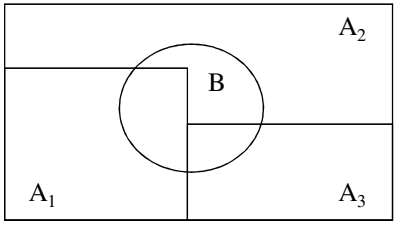
General Probability Rules

- Total Probability Rule

If $A_1,\ldots A_n$ are mutually exclusive, ***exhaustive*** events, then

$$P(B) = \sum_{i=1}^{n} P(B \cap A_i)$$

$$P(B) = \sum_{i=1}^{n} P(B \mid A_i) P(A_i)$$



General Probability Rules

- Total Probability Rule

Example

The following table gives the estimated proportion of individuals with Alzhiemer's disease by age group. It also gives the proportion of the general population that are expected to fall in the age group in 2030. What proportion of the population in 2030 will have Alzhiemer's disease?

		Proportion population	Proportion with AD	Hypoth. population	Number affected
Age group	< 65	.80	.00	80,000	0
	65 – 75	.11	.03	11,000	330
	75 – 85	.07	.11	7,000	770
	> 85	.02	.30	2,000	600
				100,000	1700

P(AD) = 0*.8 + .03*.11 + .11*.07 + .30*.02 = .017

Bayes' Rule

Bayes' rule combines multiplication rule with total probability rule

$$P(A_j \mid B) = \frac{P(A_j \cap B)}{P(B)}$$

$$= \frac{P(B \mid A_j)P(A_j)}{P(B)}$$

$$= \frac{P(B \mid A_j)P(A_j)}{\sum_{i=1}^{n} P(B \mid A_i)P(A_i)}$$

We will only apply this to the situation where A and B have two levels each, say, A and $\overline{A}$, B and $\overline{B}$. The formula becomes

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B \mid A)P(A) + P(B \mid \overline{A})P(\overline{A})}$$

Screening - An Application of Bayes' Rule

Suppose we have a random sample of a population...

		Disease Status		
		Pos.	Neg.	
Test Result	Pos.	90	30	120
	Neg.	10	970	980
		100	1000	1100

A = disease pos.
B = test pos.

Prevalence = P(A) = 100/1100 = .091

Sensitivity = P(B | A) = 90/100 = .9

Specificity = $P(\overline{B} \mid \overline{A})$ = 970/1000 = .97

PVP = P(A | B) = 90/120 = .75

PVN = $P(\overline{A} \mid \overline{B})$ = 970/980 = .99

Screening - An Application of Bayes' Rule

Now suppose we have taken a sample of 100 disease positive and 100 disease negative individuals (e.g. case-control design)

		Disease Status		
		Pos.	Neg.	
Test Result	Pos.	90	3	93
	Neg.	10	97	107
		100	100	200

A = disease pos.
B = test pos.

Prevalence = ???? (not .5!)

Sensitivity = $P(B \mid A) = 90/100 = .9$

Specificity = $P(\overline{B} \mid \overline{A}) = 97/100 = .97$

PVP = $P(A \mid B) = 90/93$ **NO!**

PVN = $P(\overline{A} \mid \overline{B}) = 97/107$ **NO!**

Screening - An Application of Bayes Rule

A = disease pos.
B = test pos.

Assume we know, from external sources, that P(A) = 100/1100. Then for every 100 disease positives we should have 1000 disease negatives …. 1:10.

Make a mock table …

		Disease Status		
		Pos.	Neg.	
Test Result	Pos.	90	3×10	120
	Neg.	10	97×10	980
		100	100×10	1100

$$PVP = \frac{90}{90 + 3 \times 10} = .75$$

Screening - an application of Bayes Rule

Now, use Bayes rule …

$$\begin{aligned} PVP = P(A|B) &= \frac{P(B|A)P(A)}{P(B)} \\ &= \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|\bar{A})P(\bar{A})} \\ &= \frac{.9 \times {}^{100}/_{1100}}{.9 \times {}^{100}/_{1100} + .03 \times {}^{1000}/_{1100}} \\ &= \frac{.9 \times 100}{.9 \times 100 + .03 \times 1000} = .75 \end{aligned}$$

Summary

- Probability - meaning
  - 1) classical
  - 2) frequentist
  - 3) subjective (personal)
- Sample space, events
- Mutually exclusive, independence
- and, or, complement
- Joint, marginal, conditional probability
- Probability - rules
  - 1) Addition
  - 2) Multiplication
  - 3) Total probability
  - 4) Bayes
- Screening
  - •sensitivity
  - •specificity
  - •predictive values

Problems

1. If allelle A has frequency 3/4 and allelle a has frequency 1/4 , what are the prevalences of the 3 genotypes AA, Aa and aa in the population (assuming random mating)?
2. A certain operation has a survival rate of 70%. If this operation is performed independently on three different patients, what is the probability all three operations will fail?
3. Suppose an influenza epidemic strikes a city. In 10% of (two parent) families the mother has influenza (event A); in 10% of families the father has influenza (event B) and in 1% of families both the mother and father have influenza.
   a) Are the events A and B independent?
   b) What is the probability neither the mother nor father have influenza?
4. The following table gives the probability of disease for different alleles of a gene (penetrances). What is the predcited probability of disease on a randomly selected individual if you have no genetic information? (Hint: use the total probability rule)

Allele	Proportion with this allele	Probability of disease with this allele
A1	.0004	.540
A2	.0059	.813
A3	.0855	.379
A4	.9082	0.0

5. In a group of symptomatic women attending a clinic, some had cervical infections with *Chlamydia trachomatis* (C) or *Neisseria gonorrhea* (G), and some were harboring both organisms. Seven women had C only, 5 women had G only and 8 women had both (B).
   a) What is the probability of any chlamydia (C) present?
   b) What is the probability of any gonorrhea (G) present?
   c) What is the probability of any gonorrhea (G) or chlamydia (C) present?
   d) Are gonorrhea and chlamydia mutually exclusive?

Problems

6) The following table summarizes a famous study by Jerushalmy et al that sparked controversy concerning the value of various screening procedures for disease detection.

	Persons without TB	Persons with TB	Total
Negative X-ray	1739	8	1747
Positive X-ray	51	22	73
Total	1790	30	1820

a) If one of the 1820 records were randomly selected, what is the probability it would be a person with TB?
b) For a randomly selected record, what is the probability that it belongs to a person who has TB and has a positive X-ray?
c) If you are told that a randomly selected record is for a person with a positive X-ray, what is the probability that it belongs to a person with TB?
d) What is the probability that a randomly selected record belongs to a person with TB or a person with a positive X-ray?

7) Estimates of the proportion of individuals with Alzheimer's disease (AD) in various age and gender groups is given in the following table. Suppose an unrelated 77 year old man, 76 year old woman and 82 year old woman are selected from the community represented in this sample. Each will be tested for AD.

Age group	Males	Females
65-69	0.016	0.0
70-74	0.0	0.022
75-79	0.049	0.023
80-84	0.086	0.078
85+	0.35	0.279

a)The sample space for this "experiment" consists of all possible outcomes of the testing. List these (hint: there are 8 possible outcomes).
b)What is the probability all three have AD?
c)What is the probability at least one has AD?
d)What is the probability exactly one has AD?

Solutions

1) P(AA) = (3/4)*(3/4)= 9/16 P(Aa) = 2*3/16 = 6/16 P(aa) = 1/16
2) P(fail,fail,fail) = P(fail)P(fail)P(fail) = .3*.3*.3 = .027
3) a) Yes, since .1*.1 = .01

   b) P(neither) = .9*.9 = .81
4) Prob = .0004*.54 + .0059*.813 + .0855*.379 + .9082*0 = .037
5) a) (7+8)/20 = .75

   b) (5+8)/20 = .65

   c) 20/20 = 1

   d) No, can have both
6) a) 30/1820

   b) 22/1820

   c) 22/73

   d) (30+73-22)/1820
7) a) Let A = has AD; a = does not have AD

77yo	76yo	82yo	Prob
A	A	A	.049*.023*.078
A	A	a	.049*.023*(1-.078)
A	a	A	etc
A	a	a	
a	A	A	
a	A	a	
a	a	A	
a	a	a	

   b) .049*.023*.078 = 8.7906e-05

   c) 1 – P(aaa) = 1 – (1-.049)(1-.023)(1-.078) = .143

   d) P(Aaa)+P(aAa)+P(aaA) = .136

Descriptive Statistics and Exploratory Data Analysis

Descriptive Statistics (Exploratory)

- "Exploratory data analysis is detective work - numerical detective work"
- "Exploratory data analysis can never be the whole story, but nothing else can serve as the foundation stone- the first step"

  John Tukey
  Exploratory Data Analysis
  Addison-Wesley, 1977

- organization, summarization, and presentation of data
- If you can't see it, don't believe it!

Inferential Statistics (Confirmatory)

- Generalization of conclusions:

  sample $\longrightarrow$ population
- Assess strength of evidence
- Make comparisons
- Make predictions

Tools:

- Modeling
- Estimation and Confidence Intervals
- Hypothesis Testing

Exploratory vs Confirmatory Data Analysis

Exploratory (Descriptive)

- Detective work
- Open (but directed) mind
- Creative

Confirmatory (Inferential)

- Acting as judge and jury (or at least lawyer)
- Focused on one or a few ideas
- Following principles of inference

Types of Data

- Categorical (qualitative)
  - 1) Nominal scale - no natural order
    - gender, marital status, race
  - 2) Ordinal scale
    - severity scale, good/better/best
- Numerical (quantitative)
  - 1) Discrete - (few) integer values
    - number of children in a family
  - 2) Continuous - measure to arbitrary precision
    - blood pressure, weight

Why bother?

$\Rightarrow$ PROPER DISPLAYS

$\Rightarrow$ PROPER ANALYSIS

Samples

In statistics we usually deal with a **sample** of observations or measurements. We will denote a sample of N numerical values as:

$$X_1, X_2, X_3, \ldots, X_N$$

where X_1 is the first sampled datum, X_2 is the second, etc.

Sometimes it is useful to order the measurements. We denote the ordered sample as:

$$X_{(1)}, X_{(2)}, X_{(3)}, \ldots, X_{(N)}$$

where $X_{(1)}$ is the smallest value and $X_{(N)}$ is the largest.

Arithmetic Mean

The **arithmetic mean** is the most common measure of the **central location** of a sample. We use $\overline{X}$ to refer to the mean and define it as:

$$\overline{X} = \frac{1}{N}\sum_{i=1}^{N} X_i$$

The symbol Σ is shorthand for "*sum*" over a specified range. For example:

$$\sum_{i=1}^{4} X_i = (X_1 + X_2 + X_3 + X_4)$$

Some Properties of the Arithmetic Mean

Often we wish to **transform** variables. Linear changes to variables (i.e. Y = a*X+b) impact the mean in a predictable way:

(1) Adding (or subtracting) a constant to all values:

$$Y_i = X_i + c$$
$$\overline{Y} =$$

(2) Multiplication (or division) by a constant:

$$Y_i = cX_i$$
$$\overline{Y} =$$

Does this nice behavior happen for any change? NO! (show that $\log \overline{X} \neq \overline{\log X}$)

Median

Another measure of central tendency is the **median** - the "middle one". Half the values are below the median and half are above. Given the ordered sample, $X_{(i)}$, the median is:

N odd:

$$\text{Median} = X_{\left(\frac{N+1}{2}\right)}$$

N even:

$$\text{Median} = \frac{1}{2}\left(X_{(N/2)} + X_{(N/2+1)}\right)$$

Mode

The **mode** is the most frequently occurring value in the sample.

Comparison of Mean and Median

- Mean is sensitive to a few very large (or small) values - "outliers"
- Median is "resistant" to outliers
- Mean is attractive mathematically
- 50% of sample is above the median,

  50% of sample is below the median.

Variation is important!

Measures of Spread: Range

The **range** is the difference between the largest and smallest observations:

$$\begin{aligned} \text{Range} &= \text{Maximum} - \text{Minimum} \\ &= X_{(N)} - X_{(1)} \end{aligned}$$

Alternatively, the range may be denoted as the pair of observations:

$$\begin{aligned} \text{Range} &= (\text{Minimum}, \text{Maximum}) \\ &= \left(X_{(1)}, X_{(N)}\right) \end{aligned}$$

The latter form is useful for data quality control..

Disadvantage: the sample range increases with increasing sample size.

Measures of Spread: Variance

Consider the following two samples:

20,23,34,26,30,22,40,38,37

30,29,30,31,32,30,28,30,30

These samples have the same mean and median, but the second is much less variable. The average "distance" from the center is quite small in the second. We use the **variance** to describe this feature:

$$s^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(X_i - \overline{X}\right)^2$$

$$s^2 = \frac{1}{N-1}\left(\sum_{i=1}^{N} X_i^2 - N\overline{X}^2\right)$$

$$s^2 = \frac{1}{N-1}\left(\sum_{i=1}^{N} X_i^2 - (\sum_{i=1}^{N} X_i)^2 / N\right)$$

The standard deviation is simply the square root of the variance:

$$\text{standard deviation} = s = \sqrt{s^2}$$

For the first sample, we obtain:

$$\overline{X} = 30$$

$$\sum_{i=1}^{9} X_i^2 = 8574$$

$$s^2 = \frac{1}{9-1}\left(8574 - 9\times 30^2\right) = (8574 - 8100)/8 = 59.25\text{yr}^2$$

For the second sample, we obtain:

$$\overline{X} = 30$$

$$\sum_{i=1}^{9} X_i^2 = 8110$$

$$s^2 = \frac{1}{9-1}\left(8110 - 9\times 30^2\right) = (8110 - 8100)/8 = 1.25\text{yr}^2$$

Properties of the variance/standard deviation

- Variance and standard deviation are **ALWAYS** greater than or equal to zero.
- Linear changes are a little trickier than they were for the mean:

(1) Add/substract a constant: $Y_i=X_i+c$

$$\begin{aligned} s_Y^2 &= \frac{1}{N-1}\sum_{i=1}^{N}\left(Y_i-\overline{Y}\right)^2 \\ &= \frac{1}{N-1}\sum_{i=1}^{N}\left(X_i+c-(\overline{X}+c)\right)^2 \\ &= s_X^2 \end{aligned}$$

(2) Multiply/divide by a constant: $Y_i=c \times X_i$

$$\begin{aligned} s_Y^2 &= \frac{1}{N-1}\sum_{i=1}^{N}\left(Y_i-\overline{Y}\right)^2 \\ &= \frac{1}{N-1}\sum_{i=1}^{N}\left(cX_i-c\overline{X}\right)^2 \\ &= \frac{1}{N-1}\sum_{i=1}^{N}c^2\left(X_i-\overline{X}\right)^2 \\ &= c^2 \times s_X^2 \end{aligned}$$

So what happens to the standard deviation?

Measures of Spread: Quantiles and Percentiles

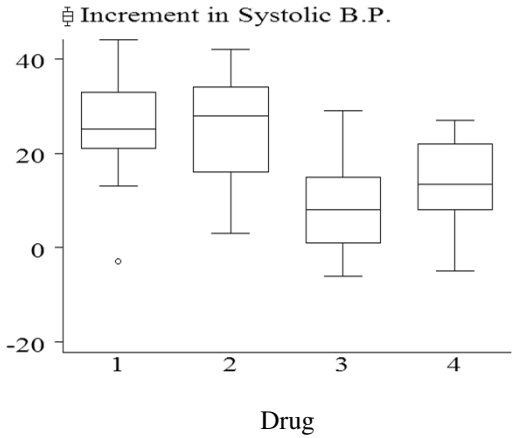
The median was the sample value that had 50% of the data below it.

More generally, we define the **p^{th} percentile** as the value which has p% of the sample values less than or equal to it.

Quartiles are the (25,50,75) percentiles. The **interquartile range** is $Q_{.75}$-$Q_{.25}$ and is another useful measure of spread. The middle 50% of the data is found between $Q_{.25}$ and is $Q_{.75}$.

Boxplot

A graphics display of the quartiles of a dataset, as well as the range. Extremely large or small values are also identified.



Summary

- Numerical Summaries
  1. location - mean, median, mode.
  2. spread - range, variance, standard deviation, IQR
- Graphical Summaries
  1. Boxplot

Probability Distributions

I

Probability Distribution

Definition: A **random variable** is a characteristic whose obtained values arise as a result of chance factors.

Definition: A **probability distribution** gives the probability of obtaining all possible (sets of) values of a random variable. It gives the probability of the outcomes of an experiment. Note that a probability distribution is an example of the "classical" definition of probability.

Population	⟷ Sample
Random variable	⟷ Measurement
Probability dist.	⟷ Frequency dist.
Parameters	⟷ Statistics (Estimates)

Theoretical Distributions

Used to provide a mathematical description of outcomes of an experiment.

A. Discrete variables

1. Binomial - sums of 0/1 outcomes

   - underlies many epidemiologic applications

   - basic model for logistic regression

2. Multinomial – generalization of binomial

   - a basic model for log-linear analysis

B. Continuous variables

1. Normal - bell-shaped curve; many measurements are approximately normally distributed.

2. t- distribution

3. Chi-square distribution (χ^2)

Binomial Distribution - Motivation

Question: In a family where both parents are carriers for a recessive trait, what is the probability that in a family of 3 children exactly 1 child would be affected?

What is the probability that at least 1 would be affected?

In a family of 6 children, what is the probability that exactly 1 child is affected?

What if the trait is dominant?

Bernoulli Trial

A Bernoulli trial is an experiment with only 2 possible outcomes, which we denote by 0 or 1 (e.g. coin toss)

Assumptions:

1) Two possible outcomes - success (1) or failure (0).

2) The probability of success, p, is the same for each trial.

3) The outcome of one trial has no influence on later outcomes (independent trials).

Binomial Random Variable

A binomial random variable is simply the total number of successes in **n** Bernoulli trials.

Example: number of affected children in a family of 3.

What we need to know is:

1. How many ways are there to get k successes (k=0,…3) in n trials?

2. What's the probability of any given outcome with exactly k successes (does order matter)?

Combinations

Combinations: number of different arrangements of k objects (successes) taken from a total of n objects (trials) if order doesn't matter.

$$C_k^n = \binom{n}{k} = \frac{n!}{k!(n-k)!}$$

"n factorial" = n! = n × (n-1) × … × 1

E.g.

Child number 1	2	3	Outcomes
+	+	+	3 affected
+	+	-	2 affected
+	-	+	2 affected
-	+	+	2 affected
+	-	-	1 affected
-	+	-	1 affected
-	-	+	1 affected
-	-	-	0 affected

What are the probabilities of these outcomes?

Child number 1	2	3	Outcomes	# ways
p	p	p	3 affected	1
p	p	1-p	2 affected	
p	1-p	p	2 affected	3
1-p	p	p	2 affected	
p	1-p	1-p	1 affected	
1-p	p	1-p	1 affected	3
1-p	1-p	p	1 affected	
1-p	1-p	1-p	0 affected	1

sequence of k +'s (0, 1, 2, or 3) and (3-k) –'s will have probability

$$p^k(1-p)^{3-k}$$

But there are $\frac{3!}{k!(3-k)!}$ such sequences, so in general…

Binomial Probabilities

What is the probability that a binomial random variable with **n** trials and success probability **p** will yield exactly **k** successes?

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

This formula is called the **probability mass function** for the binomial distribution.

Assumptions:

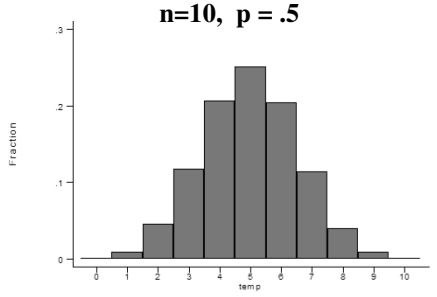
1) Two possible outcomes - success (1) or failure (0) - for each of n trials.

2) The probability of success, p, is the same for each trial.

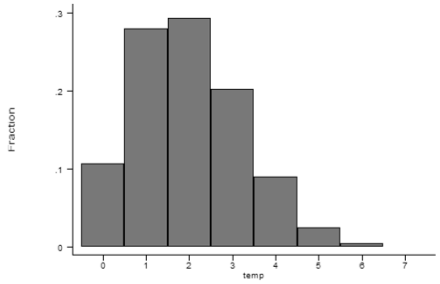
3) The outcome of one trial has no influence on later outcomes (independent trials).

4) The random variable of interest is the total number of successes.

n=10, p = .5



n=10, p = .2



Binomial Probabilities - Example

Returning to the original question: What is the probability of exactly 1 affected child in a family of 3? (recessive trait, carrier parents)

Mean and Variance of a Discrete Random Variable

Given a **theoretical** probability distribution we can define the **mean and variance of a random variable** which follows that distribution. These concepts are analogous to the summary measures used for samples except that these now describe the value of these summaries in the limit as the sample size goes to infinity (i.e. the **parameters** of the **population**).

Suppose a random variable X can take the values $\{x_1,x_2,\ldots\}$ with probabilities $\{p_1,p_2,\ldots\}$. Then

MEAN:

$$\mu = \mathrm{E}(\mathrm{X}) = \sum_{\mathrm{j}} p_j x_j$$

VARIANCE:

$$\sigma^2 = \mathrm{V}(\mathrm{X}) = \mathrm{E}\left[(\mathrm{X} - \mu)^2\right] = \sum_{j} p_j (x_j - \mu)^2$$

Example - Mean and Variance

Consider a Bernoulli random variable with success probability **p**.

P[X=1] = p

P[X=0]=1-p

MEAN:

$$\begin{aligned}\mu = E[X] &= \sum_{j=0}^{1} p_j x_j \\ &= (1-p)\times 0 \quad + \quad p\times 1 \\ &= p\end{aligned}$$

VARIANCE

$$\begin{aligned}\sigma^2 = \mathrm{V}[X] &= \sum_{j=0}^{1} p_j (x_j - \mu)^2 \\ &= (1-p)\times(0-p)^2 \quad + \quad p\times(1-p)^2 \\ &= p(1-p)\end{aligned}$$

Mean and Variance - Binomial

Consider a binomial random variable with success probability **p** and sample size **n**.

X ~ bin(n,p)

MEAN:

$$\begin{aligned}\mu = E[X] &= \sum_{j=0}^{n} p_j x_j \\ &= \sum_{j=0}^{n} \binom{n}{j} p^j (1-p)^{n-j} \times j \\ &= ???\end{aligned}$$

VARIANCE:

$$\begin{aligned}\sigma^2 = \mathrm{V}[X] &= \sum_{j=0}^{n} p_j (x_j - \mu)^2 \\ &= \sum_{j=0}^{n} \binom{n}{j} p^j (1-p)^{n-j} \times (j-\mu)^2 \\ &= ???\end{aligned}$$

Help!

Means and Variance of the Sum of independent RV's

Recall that a binomial RV is just the **sum** of **n** independent Bernoulli random variables.

If $X_1, X_2, \ldots, X_n$ are **independent** random variables and if we define $Y = X_1 + X_2 + \ldots + X_n$

1. Means add:

$$E[Y] = E[X_1] + E[X_2] + \ldots + E[X_n]$$

2. Variances add:

$$V[Y] = V[X_1] + V[X_2] + \ldots + V[X_n]$$

We can use these results, together with the properties of the mean and variance that we learned earlier, to obtain the mean and variance of a binomial random variable (hmwk).

Binomial Distribution Summary

Binomial

1. Discrete, bounded
2. Parameters - **n,p**
3. Sum of n independent 0/1 outcomes
4. Sample proportions, logistic regression

Probability Distributions

II

Multinomial Distribution - Motivation

Suppose we modified assumption (1) of the binomial distribution to allow for more than two outcomes.

For example, suppose that for the family with parents that are heterozygote carriers of a recessive trait, we are interested in knowing the probability of

Q_1: One of their n=3 offspring will be unaffected (AA), 1 will be affected (aa) and one will be a carrier (Aa),

Q_2:All of their offspring will be carriers,

Q_3:Exactly two of their offspring will be affected (aa) and one will be a carrier.

Multinomial Distribution - Motivation

For each child, we can represent these possibilities with three indicator variables for the i-th child as

Y_{i1} = 1 if unaffected (AA), & 0 otherwise
Y_{i2} = 1 if carrier (Aa), & 0 otherwise
Y_{i3} = 1 if affected (aa), & 0 otherwise

Notice only one of the three Y_{i1}, Y_{i2}, Y_{i3} can be equal to 1, so $\Sigma_j Y_{ij} = 1$.

For the binomial distribution with 2 outcomes, there are 2^n unique outcomes in n trials. In the family with n=3 children, there are $2^3 = 8$ unique outcomes.

For the multinomial distribution with n trials and only 3 outcomes, the number of unique outcomes is 3^n. For our small family, that's 3^3=27 outcomes.

Possible Outcomes

Combinations: As with the binomial, there are different ways to arrange possible outcomes from a total of n objects (trials) if order doesn't matter. For the multinomial distribution, the combinations are summarized as

$$C_k^n = \frac{n!}{k_1! k_2 \cdots k_J!}$$

where the k_j (j=1,2,…,J) correspond to the totals for the different outcomes.

E.g. (n=2 offspring)

Child number 1	2	Outcomes
AA	AA	2 unaffected, 0 carrier, 0 affected
AA	Aa	1 unaffected, 1 carrier, 0 affected
Aa	AA	1 unaffected, 1 carrier, 0 affected
AA	aa	1 unaffected, 0 carrier, 1 affected
aa	AA	1 unaffected, 0 carrier, 1affected
Aa	Aa	0 unaffected, 2 carrier, 0 affected
aa	Aa	0 unaffected, 1 carrier, 1 affected
Aa	aa	0 unaffected, 1 carrier, 1 affected
aa	aa	0 unaffected, 0 carrier, 2 affected

For the case of n=2 offspring (i.e., trials), what are the probabilities of these outcomes?

E.g. (n=2, k_1=unaffected, k_2=carrier, k_3=affected)

Child number 1	2	Outcomes	# ways
p_1	p_1	k_1=2,k_2=0,k_3=0	1
p_1	p_2	k_1=1,k_2=1,k_3=0	2
p_2	p_1	k_1=1,k_2=1,k_3=0	
p_1	p_3	k_1=1,k_2=0,k_3=1	2
p_3	p_1	k_1=1,k_2=0,k_3=1	
p_2	p_2	k_1=0,k_2=2,k_3=0	1
p_3	p_2	k_1=0,k_2=1,k_3=1	2
p_2	p_3	k_1=0,k_2=1,k_3=1	
p_3	p_3	k_1=0,k_2=0,k_3=2	1

For each possible outcome, the probability $\Pr[Y_1=k_1, Y_2=k_2, Y_3=k_3]$ is

$$p_1^{k1}p_2^{k2}p_3^{k3}$$

There are $\frac{n!}{k_j!}$ sequences for each probability, so in general…

Multinomial Probabilities

What is the probability that a multinomial random variable with **n** trials and success probabilities p_1, p_2, …, p_J will yield exactly k_1, k_2,…k_J successes?

$$P(Y_1=k_1, Y_2=k_2, \cdots Y_J=k_J)=\frac{n!}{k_1!k_2!\cdots k_J!}p_1^{k_1}p_2^{k_2}\cdots p_J^{k_J}$$

Assumptions:

1) J possible outcomes – only one of which can be a success (1) a given trial.

2) The probability of success for each possible outcome, p_j, is the same from trial to trial.

3) The outcome of one trial has no influence on other trials (independent trials).

4) Interest is in the (sum) total number of "successes" over all the trials.

k_1	k_2	k_3	k_4	· · ·	k_{J-1}	k_J

$n = \Sigma_j k_j$ is the total number of trials.

Multinomial Random Variable

A multinomial random variable is simply the total number of successes in n trials.

Example: family of 3 offspring.

	child 1		child 2		child 3		Total
$\mathbf{Q_1}$:	1 0 0	+	0 0 1	+	0 1 0	=	1 1 1
$\mathbf{Q_2}$:	0 1 0	+	0 1 0	+	0 1 0	=	0 3 0
$\mathbf{Q_3}$:	0 0 1	+	0 1 0	+	0 0 1	=	0 1 2

Multinomial Probabilities - Examples

Returning to the original questions:

$\mathbf{Q_1}$: One of n=3 offspring will be unaffected (AA), one will be affected (aa) and one will be a carrier (Aa) (recessive trait, carrier parents)?

Solution: For a given child, the probabilities of the three outcomes are:

p_1 = Pr[AA] = 1/4,
p_2 = Pr[Aa] = 1/2,
p_3 = Pr[aa] = 1/4.

We have

$$\begin{aligned} P(Y_1 = 1, Y_2 = 1, \cdots Y_3 = 1) &= \frac{3!}{1!1!1!} p_1^1 p_2^1 p_3^1 \\ &= \frac{(3)(2)(1)}{(1)(1)(1)} \left(\frac{1}{4}\right)^1 \left(\frac{1}{2}\right)^1 \left(\frac{1}{4}\right)^1 \\ &= \frac{3}{16} = 0.1875. \end{aligned}$$

Binomial Probabilities - Examples

Q_2: What is the probability that all three offspring will be carriers?

$$P(Y_1 = 0, Y_2 = 3, Y_3 = 0) = \frac{3!}{0!3!0!} p_1^0 p_2^3 p_3^0$$
$$= \frac{(3)(2)(1)}{(3)(2)(1)} \left(\frac{1}{4}\right)^0 \left(\frac{1}{2}\right)^3 \left(\frac{1}{4}\right)^0$$
$$= \frac{1}{8} = 0.125.$$

Q_3: What is the probability that exactly two offspring will be affected and one a carrier?

$$P(Y_1 = 0, Y_2 = 1, Y_3 = 2) = \frac{3!}{0!1!2!} p_1^0 p_2^1 p_3^2$$
$$= \frac{(3)(2)(1)}{(2)(1)} \left(\frac{1}{4}\right)^0 \left(\frac{1}{2}\right)^1 \left(\frac{1}{4}\right)^2$$
$$= \frac{3}{32} = 0.09375.$$

Example - Mean and Variance

It turns out that the (marginal) outcomes of the multinomial distribution are binomial. We can immediately obtain the means for each outcome (i.e., the i^{th} cell)

MEAN: $$E[k_j] = E\left[\sum_{i=1}^{n} Y_{ij}\right] = \sum_{i=1}^{n} E[Y_{ij}]$$
$$= \sum_{i=1}^{n} p_j = np_j$$

VARIANCE:

$$V[k_j] = V\left[\sum_{i=1}^{n} Y_{ij}\right] = \sum_{i=1}^{n} V[Y_{ij}]$$
$$= \sum_{i=1}^{n} p_j(1 - p_j) = np_j(1 - p_j)$$

COVARIANCE:

$$Cov[k_j, k_{j'}] = -np_j p_{j'}$$

Multinomial Distribution Summary

Multinomial

1. Discrete, bounded
2. Parameters - $n, p_1, p_2,\ldots,p_J$
3. Sum of n independent outcomes
4. Extends binomial distribution
5. Polytomous regression, contingency tables

Continuous Distributions

Continuous Distributions

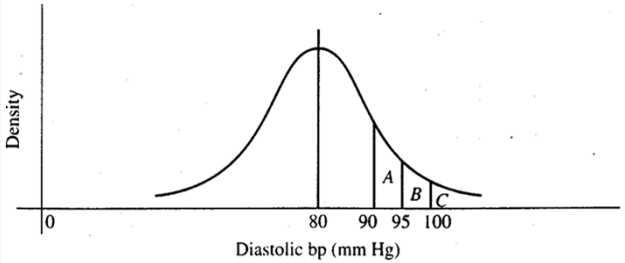
For measurements like height and weight which can be measured with arbitrary precision, it does not make sense to talk about the probability of any single value. Instead we talk about the probability for an **interval**.

$$P[\text{weight} = 70.000\text{kg}] \approx 0$$

$$P[69.0\text{kg} \leq \text{weight} \leq 71.0\text{kg}] = 0.08$$

For discrete random variables we had a probability mass function to give us the probability of each possible value. For continuous random variables we use a **probability density function** to tell us about the probability of obtaining a value within some interval.

E.g. Rosner - diastolic blood pressure in 35-44 year-old men (figure 5.1)



For any interval, the **area** under the curve represents the probability of obtaining a value in that interval.

Probability density function

1. A function, typically denoted f(x), that gives probabilities based on the **area** under the curve.
2. $f(x) \geq 0$
3. Total area under the function f(x) is 1.0.

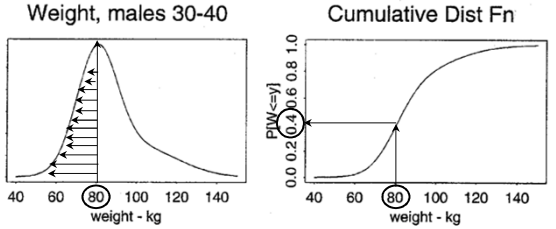
$$\int f(x)dx = 1.0$$

Cumulative distribution function

The cumulative distribution function, F(t), tells us the total probability less than some value t.

$$F(t) = P(X \leq t)$$

This is analogous to the cumulative relative frequency.



Prob[wgt < 80] = 0.40

Area under the curve

THE
NORMAL
LAW OF ERROR
STANDS OUT IN THE
EXPERIENCE OF MANKIND
AS ONE OF THE BROADEST
GENERALIZATIONS OF NATURAL
PHILOSOPHY ◆ IT SERVES AS THE
GUIDING INSTRUMENT IN RESEARCHES
IN THE PHYSICAL AND SOCIAL SCIENCES AND
IN MEDICINE AGRICULTURE AND ENGINEERING ◆
IT IS AN INDISPENSABLE TOOL FOR THE ANALYSIS AND THE
INTERPRETATION OF THE BASIC DATA OBTAINED BY OBSERVATION AND EXPERIMENT

Normal Distribution

- A common probability model for continuous data
- Can be used to characterize the Binomial or Poisson under certain circumstances
- Bell-shaped curve

$\Rightarrow$ takes values between $-\infty$ and $+\infty$

$\Rightarrow$ symmetric about mean

$\Rightarrow$ mean=median=mode

- Examples

  birthweights

  blood pressure

  CD4 cell counts (perhaps transformed)

Normal Distribution

Specifying the mean and variance of a normal distribution completely determines the probability distribution function and, therefore, all probabilities.

The **normal probability density function** is:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2} \frac{(x-\mu)^2}{\sigma^2} \right)$$

where

$\pi \approx 3.14$ (a constant)

Notice that the normal distribution has two parameters:

μ = the mean of X

σ = the standard deviation of X

We write $X \sim N(\mu, \sigma^2)$. The **standard normal** distribution is a special case where $\mu = 0$ and $\sigma = 1$.

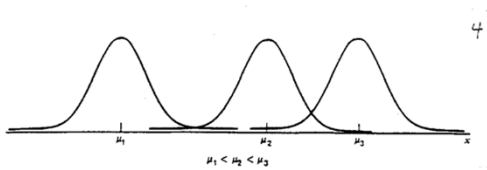


FIGURE 3.6.3
Three Normal Distributions with Different Means

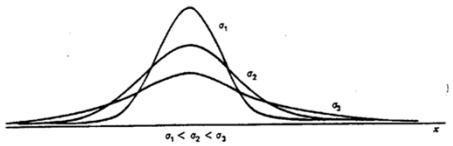


FIGURE 3.6.4
Three Normal Distributions with Different Standard Deviations

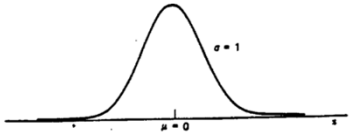
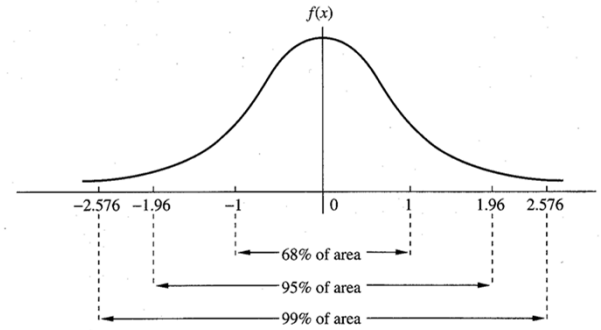


FIGURE 3.6.5
The Unit Normal Distribution



Normal Distribution - Calculating Probabilities

Example: Rosner 5.20

Serum cholesterol is approximately normally distributed with mean 219 mg/mL and standard deviation 50 mg/mL. If the clinically desirable range is < 200 mg/mL, then what proportion of the population falls in this range?

X = serum cholesterol in an individual.

μ =

σ =

$$P[x < 200] = \int_{-\infty}^{200} \frac{1}{50\sqrt{2\pi}} \exp\left(-\frac{1}{2} \frac{(x-219)^2}{50^2} \right) dx$$

negative values for cholesterol - huh?

Standard Normal Distribution - Calculating Probabilities

First, let's consider the **standard normal** - N(0,1). We will usually use Z to denote a random variable with a standard normal distribution. The density of Z is

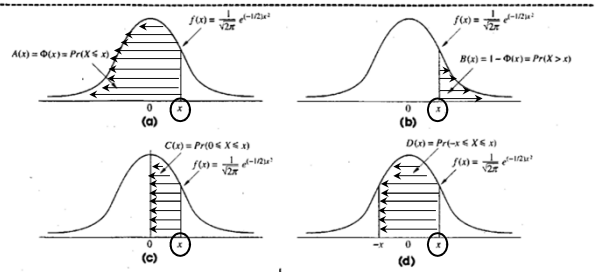
$$f(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}z^2\right)$$

and the **cumulative distribution** of Z is:

$$P(Z \le x) = \Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}z^2\right) dz$$

Tables (Rosner table 3) and computer routines are available for calculating these probabilities.

TABLE 3 The normal distribution



x	A[a]	B[b]	C[c]	D[d]	x	A	B	C	D
0.0	.5000	.5000	.0	.0	0.32	.6255	.3745	.1255	.2510
0.01	.5040	.4960	.0040	.0080	0.33	.6293	.3707	.1293	.2586
0.02	.5080	.4920	.0080	.0160	0.34	.6331	.3669	.1331	.2661
0.03	.5120	.4880	.0120	.0239	0.35	.6368	.3632	.1368	.2737
0.04	.5160	.4840	.0160	.0319	0.36	.6406	.3594	.1406	.2812
0.05	.5199	.4801	.0199	.0399	0.37	.6443	.3557	.1443	.2886
0.06	.5239	.4761	.0239	.0478	0.38	.6480	.3520	.1480	.2961
0.07	.5279	.4721	.0279	.0558	0.39	.6517	.3483	.1517	.3035
0.08	.5319	.4681	.0319	.0638	0.40	.6554	.3446	.1554	.3108
0.09	.5359	.4641	.0359	.0717	0.41	.6591	.3409	.1591	.3182
0.10	.5398	.4602	.0398	.0797	0.42	.6628	.3372	.1628	.3255
0.11	.5438	.4562	.0438	.0876	0.43	.6664	.3336	.1664	.3328
0.12	.5478	.4522	.0478	.0955	0.44	.6700	.3300	.1700	.3401
0.13	.5517	.4483	.0517	.1034	0.45	.6736	.3264	.1736	.3473
0.14	.5557	.4443	.0557	.1113	0.46	.6772	.3228	.1772	.3545
0.15	.5596	.4404	.0596	.1192	0.47	.6808	.3192	1808	.3616
0.16	.5636	.4364	.0636	.1271	0.48	.6844	.3156	1844	.3688
0.17	.5675	.4325	.0675	.1350	0.49	.6879	.3121	1879	.3759
0.18	.5714	.4286	.0714	.1428	0.50	.6915	.3085	.1915	.3829
0.19	.5753	.4247	.0753	.1507	0.51	.6950	.3050	.1950	.3899
0.20	.5793	.4207	.0793	.1585	0.52	.6985	.3015	.1985	.3969
0.21	.5832	.4168	.0832	.1663	0.53	.7019	.2981	.2019	.4039
0.22	.5871	.4129	.0871	.1741	0.54	.7054	.2946	.2054	.4108
0.23	.5910	.4090	.0910	.1819	0.55	.7088	.2912	.2088	.4177
0.24	.5948	.4052	.0948	.1897	0.56	.7123	.2877	.2123	.4245
0.25	.5987	.4013	.0987	.1974	0.57	.7157	.2843	.2157	.4313
0.26	.6026	.3974	.1026	.2051	0.58	.7190	.2810	.2190	.4381
0.27	.6064	.3936	.1064	.2128	0.59	.7224	.2776	.2224	.4448
0.28	.6103	.3897	.1103	.2205	0.60	.7257	.2743	.2257	.4515
0.29	.6141	.3859	.1141	.2282	0.61	.7291	.2709	.2291	.4581
0.30	.6179	.3821	.1179	.2358	0.62	.7324	.2676	.2324	.4647
0.31	.6217	.3783	.1217	.2434	0.63	.7357	.2643	.2357	.4713

TABLE 3 (*Continued*)

x	A[a]	B[b]	C[c]	D[d]	x	A	B	C	D
1.56	.9406	.0594	.4406	.8812	2.03	.9788	.0212	.4788	.9576
1.57	.9418	.0582	.4418	.8836	2.04	.9793	.0207	.4793	.9586
1.58	.9429	.0571	.4429	.8859	2.05	.9798	.0202	.4798	.9596
1.59	.9441	.0559	.4441	.8882	2.06	.9803	.0197	.4803	.9606
1.60	.9452	.0548	.4452	.8904	2.07	.9808	.0192	.4808	.9615
1.61	.9463	.0537	.4463	.8926	2.08	.9812	.0188	.4812	.9625
1.62	.9474	.0526	.4474	.8948	2.09	.9817	.0183	.4817	.9634
1.63	.9484	.0516	.4484	.8969	2.10	.9821	.0179	.4821	.9643
1.64	.9495	.0505	.4495	.8990	2.11	.9826	.0174	.4826	.9651
1.65	.9505	.0495	.4505	.9011	2.12	.9830	.0170	.4830	.9660
1.66	.9515	.0485	.4515	.9031	2.13	.9834	.0166	.4834	.9668
1.67	.9525	.0475	.4525	.9051	2.14	.9838	.0162	.4838	.9676
1.68	.9535	.0465	.4535	.9070	2.15	.9842	.0158	.4842	.9684
1.69	.9545	.0455	.4545	.9090	2.16	.9846	.0154	.4846	.9692
1.70	.9554	.0446	.4554	.9109	2.17	.9850	.0150	.4850	.9700
1.71	.9564	.0436	.4564	.9127	2.18	.9854	.0146	.4854	.9707
1.72	.9573	.0427	.4573	.9146	2.19	.9857	.0143	.4857	.9715
1.73	.9582	.0418	.4582	.9164	2.20	.9861	.0139	.4861	.9722
1.74	.9591	.0409	.4591	.9181	2.21	.9864	.0136	.4864	.9729
1.75	.9599	.0401	.4599	.9199	2.22	.9868	.0132	.4868	.9736
1.76	.9608	.0392	.4608	.9216	2.23	.9871	.0129	.4871	.9743
1.77	.9616	.0384	.4616	.9233	2.24	.9875	.0125	.4875	.9749
1.78	.9625	.0375	.4625	.9249	2.25	.9878	.0122	.4878	.9756
1.79	.9633	.0367	.4633	.9265	2.26	.9881	.0119	.4881	.9762
1.80	.9641	.0359	.4641	.9281	2.27	.9884	.0116	.4884	.9768
1.81	.9649	.0351	.4649	.9297	2.28	.9887	.0113	.4887	.9774
1.82	.9656	.0344	.4656	.9312	2.29	.9890	.0110	.4890	.9780
1.83	.9664	.0336	.4664	.9327	2.30	.9893	.0107	.4893	.9786
1.84	.9671	.0329	.4671	.9342	2.31	.9896	.0104	.4896	.9791
1.85	.9678	.0322	.4678	.9357	2.32	.9898	.0102	.4898	.9797
1.86	.9686	.0314	.4686	.9371	2.33	.9901	.0099	.4901	.9802
1.87	.9693	.0307	.4693	.9385	2.34	.9904	.0096	.4904	.9807
1.88	.9699	.0301	.4699	.9399	2.35	.9906	.0094	.4906	.9812
1.89	.9706	.0294	.4706	.9412	2.36	.9909	.0091	.4909	.9817
1.90	.9713	.0287	.4713	.9426	2.37	.9911	.0089	.4911	.9822
1.91	.9719	.0281	.4719	.9439	2.38	.9913	.0087	.4913	.9827
1.92	.9726	.0274	.4726	.9451	2.39	.9916	.0084	.4916	.9832
1.93	.9732	.0268	.4732	.9464	2.40	.9918	.0082	.4918	.9836
1.94	.9738	.0262	.4738	.9476	2.41	.9920	.0080	.4920	.9840
1.95	.9744	.0256	.4744	.9488	2.42	.9922	.0078	.4922	.9845
1.96	.9750	.0250	.4750	.9500	2.43	.9925	.0075	.4925	.9849
1.97	.9756	.0244	.4756	.9512	2.44	.9927	.0073	.4927	.9853
1.98	.9761	.0239	.4761	.9523	2.45	.9929	.0071	.4929	.9857
1.99	.9767	.0233	.4767	.9534	2.46	.9931	.0069	.4931	.9861
2.00	.9772	.0228	.4772	.9545	2.47	.9932	.0068	.4932	.9865
2.01	.9778	.0222	.4778	.9556	2.48	.9934	.0066	.4934	.9869
2.02	.9783	.0217	.4783	.9566	2.49	.9936	.0064	.4936	.9872

Standard Normal Probabilities

Using Rosner, table 3, find

➡ $P[Z \leq 1.65] =$

$P[Z \geq 0.5] =$

$P[-1.96 \leq Z \leq 1.96] =$

$P[-0.5 \leq Z \leq 2.0] =$

TABLE 3 The normal distribution

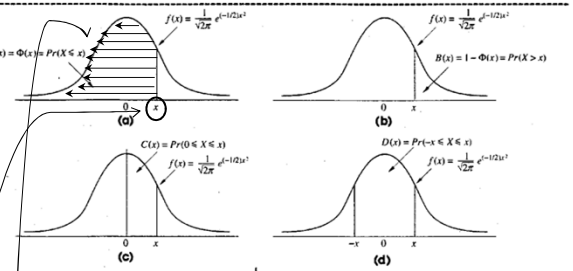


TABLE 3 (Continued)

x	A[a]	B[b]	C[c]	D[d]	x	A	B	C	D
1.56	.9406	.0594	.4406	.8812	2.03	.9788	.0212	.4788	.9576
1.57	.9418	.0582	.4418	.8836	2.04	.9793	.0207	.4793	.9586
1.58	.9429	.0571	.4429	.8859	2.05	.9798	.0202	.4798	.9596
1.59	.9441	.0559	.4441	.8882	2.06	.9803	.0197	.4803	.9606
1.60	.9452	.0548	.4452	.8904	2.07	.9808	.0192	.4808	.9615
1.61	.9463	.0537	.4463	.8926	2.08	.9812	.0188	.4812	.9625
1.62	.9474	.0526	.4474	.8948	2.09	.9817	.0183	.4817	.9634
1.63	.9484	.0516	.4484	.8969	2.10	.9821	.0179	.4821	.9643
1.64	.9495	.0505	.4495	.8990	2.11	.9826	.0174	.4826	.9651
1.65	.9505	.0495	.4505	.9011	2.12	.9830	.0170	.4830	.9660
1.66	.9515	.0485	.4515	.9031	2.13	.9834	.0166	.4834	.9668
1.67	.9525	.0475	.4525	.9051	2.14	.9838	.0162	.4838	.9676
1.68	.9535	.0465	.4535	.9070	2.15	.9842	.0158	.4842	.9684
1.69	.9545	.0455	.4545	.9090	2.16	.9846	.0154	.4846	.9692
1.70	.9554	.0446	.4554	.9109	2.17	.9850	.0150	.4850	.9700
1.71	.9564	.0436	.4564	.9127	2.18	.9854	.0146	.4854	.9707
1.72	.9573	.0427	.4573	.9146	2.19	.9857	.0143	.4857	.9715
1.73	.9582	.0418	.4582	.9164	2.20	.9861	.0139	.4861	.9722
1.74	.9591	.0409	.4591	.9181	2.21	.9864	.0136	.4864	.9729
1.75	.9599	.0401	.4599	.9199	2.22	.9868	.0132	.4868	.9736
1.76	.9608	.0392	.4608	.9216	2.23	.9871	.0129	.4871	.9743
1.77	.9616	.0384	.4616	.9233	2.24	.9875	.0125	.4875	.9749
1.78	.9625	.0375	.4625	.9249	2.25	.9878	.0122	.4878	.9756
1.79	.9633	.0367	.4633	.9265	2.26	.9881	.0119	.4881	.9762
1.80	.9641	.0359	.4641	.9281	2.27	.9884	.0116	.4884	.9768
1.81	.9649	.0351	.4649	.9297	2.28	.9887	.0113	.4887	.9774
1.82	.9656	.0344	.4656	.9312	2.29	.9890	.0110	.4890	.9780
1.83	.9664	.0336	.4664	.9327	2.30	.9893	.0107	.4893	.9786
1.84	.9671	.0329	.4671	.9342	2.31	.9896	.0104	.4896	.9791

Standard Normal Probabilities

Using Rosner, table 3, find

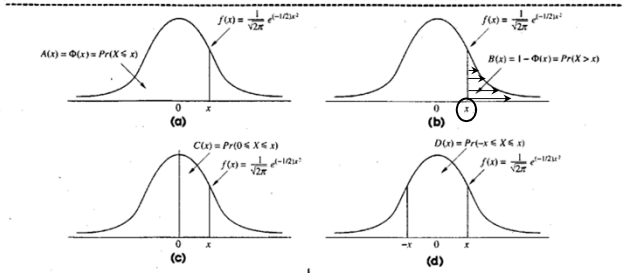
$P[Z \leq 1.65] =$ **0.9505.**

➡ $P[Z \geq 0.5] =$

$P[-1.96 \leq Z \leq 1.96] =$

$P[-0.5 \leq Z \leq 2.0] =$

TABLE 3 The normal distribution



x	A[a]	B[b]	C[c]	D[d]	x	A	B	C	D
0.0	.5000	.5000	.0	.0	0.32	.6255	.3745	.1255	.2510
0.01	.5040	.4960	.0040	.0080	0.33	.6293	.3707	.1293	.2586
0.02	.5080	.4920	.0080	.0160	0.34	.6331	.3669	.1331	.2661
0.03	.5120	.4880	.0120	.0239	0.35	.6368	.3632	.1368	.2737
0.04	.5160	.4840	.0160	.0319	0.36	.6406	.3594	.1406	.2812
0.05	.5199	.4801	.0199	.0399	0.37	.6443	.3557	.1443	.2886
0.06	.5239	.4761	.0239	.0478	0.38	.6480	.3520	.1480	.2961
0.07	.5279	.4721	.0279	.0558	0.39	.6517	.3483	.1517	.3035
0.08	.5319	.4681	.0319	.0638	0.40	.6554	.3446	.1554	.3108
0.09	.5359	.4641	.0359	.0717	0.41	.6591	.3409	.1591	.3182
0.10	.5398	.4602	.0398	.0797	0.42	.6628	.3372	.1628	.3255
0.11	.5438	.4562	.0438	.0876	0.43	.6664	.3336	.1664	.3328
0.12	.5478	.4522	.0478	.0955	0.44	.6700	.3300	.1700	.3401
0.13	.5517	.4483	.0517	.1034	0.45	.6736	.3264	.1736	.3473
0.14	.5557	.4443	.0557	.1113	0.46	.6772	.3228	.1772	.3545
0.15	.5596	.4404	.0596	.1192	0.47	.6808	.3192	1808	.3616
0.16	.5636	.4364	.0636	.1271	0.48	.6844	.3156	1844	.3688
0.17	.5675	.4325	.0675	.1350	0.49	.6879	.3121	1879	.3759
0.18	.5714	.4286	.0714	.1428	0.50	.6915	.3085	.1915	.3829
0.19	.5753	.4247	.0753	.1507	0.51	.6950	.3050	.1950	.3899
0.20	.5793	.4207	.0793	.1585	0.52	.6985	.3015	.1985	.3969
0.21	.5832	.4168	.0832	.1663	0.53	.7019	.2981	.2019	.4039
0.22	.5871	.4129	.0871	.1741	0.54	.7054	.2946	.2054	.4108
0.23	.5910	.4090	.0910	.1819	0.55	.7088	.2912	.2088	.4177
0.24	.5948	.4052	.0948	.1897	0.56	.7123	.2877	.2123	.4245
0.25	.5987	.4013	.0987	.1974	0.57	.7157	.2843	.2157	.4313
0.26	.6026	.3974	.1026	.2051	0.58	.7190	.2810	.2190	.4381
0.27	.6064	.3936	.1064	.2128	0.59	.7224	.2776	.2224	.4448
0.28	.6103	.3897	.1103	.2205	0.60	.7257	.2743	.2257	.4515
0.29	.6141	.3859	.1141	.2282	0.61	.7291	.2709	.2291	.4581
0.30	.6179	.3821	.1179	.2358	0.62	.7324	.2676	.2324	.4647
0.31	.6217	.3783	.1217	.2434	0.63	.7357	.2643	.2357	.4713

Standard Normal Probabilities

Using Rosner, table 3, find

$P[Z \leq 1.65] =$ **0.9505.**

$P[Z \geq 0.5] =$ **0.3085.**

➡ $P[-1.96 \leq Z \leq 1.96] =$

$P[-0.5 \leq Z \leq 2.0] =$

TABLE 3 The normal distribution

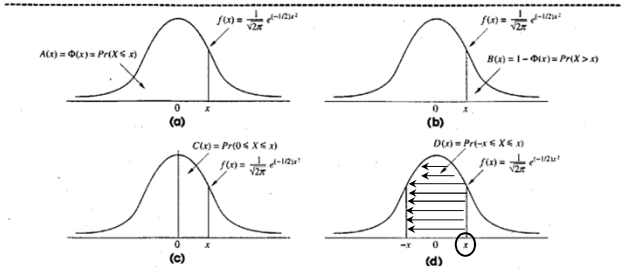


TABLE 3 (*Continued*)

x	A[a]	B[b]	C[c]	D[d]	x	A	B	C	D
1.74	.9591	.0409	.4591	.9181	2.21	.9864	.0136	.4864	.9729
1.75	.9599	.0401	.4599	.9199	2.22	.9868	.0132	.4868	.9736
1.76	.9608	.0392	.4608	.9216	2.23	.9871	.0129	.4871	.9743
1.77	.9616	.0384	.4616	.9233	2.24	.9875	.0125	.4875	.9749
1.78	.9625	.0375	.4625	.9249	2.25	.9878	.0122	.4878	.9756
1.79	.9633	.0367	.4633	.9265	2.26	.9881	.0119	.4881	.9762
1.80	.9641	.0359	.4641	.9281	2.27	.9884	.0116	.4884	.9768
1.81	.9649	.0351	.4649	.9297	2.28	.9887	.0113	.4887	.9774
1.82	.9656	.0344	.4656	.9312	2.29	.9890	.0110	.4890	.9780
1.83	.9664	.0336	.4664	.9327	2.30	.9893	.0107	.4893	.9786
1.84	.9671	.0329	.4671	.9342	2.31	.9896	.0104	.4896	.9791
1.85	.9678	.0322	.4678	.9357	2.32	.9898	.0102	.4898	.9797
1.86	.9686	.0314	.4686	.9371	2.33	.9901	.0099	.4901	.9802
1.87	.9693	.0307	.4693	.9385	2.34	.9904	.0096	.4904	.9807
1.88	.9699	.0301	.4699	.9399	2.35	.9906	.0094	.4906	.9812
1.89	.9706	.0294	.4706	.9412	2.36	.9909	.0091	.4909	.9817
1.90	.9713	.0287	.4713	.9426	2.37	.9911	.0089	.4911	.9822
1.91	.9719	.0281	.4719	.9439	2.38	.9913	.0087	.4913	.9827
1.92	.9726	.0274	.4726	.9451	2.39	.9916	.0084	.4916	.9832
1.93	.9732	.0268	.4732	.9464	2.40	.9918	.0082	.4918	.9836
1.94	.9738	.0262	.4738	.9476	2.41	.9920	.0080	.4920	.9840
1.95	.9744	.0256	.4744	.9488	2.42	.9922	.0078	.4922	.9845
1.96	.9750	.0250	.4750	.9500	2.43	.9925	.0075	.4925	.9849
1.97	.9756	.0244	.4756	.9512	2.44	.9927	.0073	.4927	.9853
1.98	.9761	.0239	.4761	.9523	2.45	.9929	.0071	.4929	.9857
1.99	.9767	.0233	.4767	.9534	2.46	.9931	.0069	.4931	.9861
2.00	.9772	.0228	.4772	.9545	2.47	.9932	.0068	.4932	.9865
2.01	.9778	.0222	.4778	.9556	2.48	.9934	.0066	.4934	.9869
2.02	.9783	.0217	.4783	.9566	2.49	.9936	.0064	.4936	.9872

Standard Normal Probabilities

Using Rosner, table 3, find

$P[Z \leq 1.65] =$ **0.9505.**

$P[Z \geq 0.5] =$ **0.3085.**

$P[-1.96 \leq Z \leq 1.96] =$ **0.9500**

➡ $P[-0.5 \leq Z \leq 2.0] =$ **?**

$= P[-0.5 \leq Z \leq 0] + P[0 \leq Z \leq 2.0]$

$= P[0 \leq Z \leq 0.5] + P[0 \leq Z \leq 2.0]$

using column (c) from Table 3

$=$ **0.1915 + 0.4772 = 0.6687.**

Converting to Standard Normal

This solves the problem for the N(0,1) case. Do we need a special table for every (μ,σ)?

No!

Define: $X = \mu + \sigma Z$ where $Z \sim N(0,1)$

1. $E(X) = \mu + \sigma E(Z) = \mu$
2. $V(X) = \sigma^2 V(Z) = \sigma^2$.
3. X is normally distributed!

Linear functions of normal RV's are also normal.

If $X \sim N(\mu, \sigma^2)$ and $Y = aX + b$ then

$$Y \sim N(a\mu + b, a^2\sigma^2)$$

Converting to Standard Normal

How can we convert a $N(\mu,\sigma^2)$ to a standard normal?

Standardize:

$$Z = \frac{X - \mu}{\sigma}$$

What is the mean and variance of Z?

1. $E(Z) = (1/\sigma)E(X - \mu) = 0$
2. $V(Z) = (1/\sigma^2)V(X) = 1$

Normal Distribution - Calculating Probabilities

Return to cholesterol example (Rosner 5.20)

Serum cholesterol is approximately normally distributed with mean 219 mg/mL and standard deviation 50 mg/mL. If the clinically desirable range is < 200 mg/mL, then what proportion of the population falls in this range?

$$
\begin{aligned}
P(X < 200) &= P\left(\frac{X-\mu}{\sigma} < \frac{200-219}{50}\right) \\
&= P\left(Z < \frac{200-219}{50}\right) \\
&= P(Z < -0.38) \\
&= P(Z > 0.38) \quad \textit{from Table 3, column (b)} \\
&= 0.3520.
\end{aligned}
$$

Normal Approximation to Binomial

Example

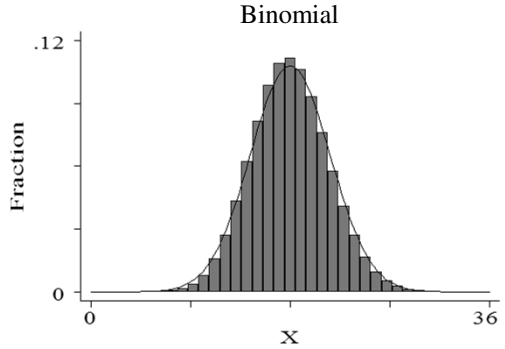
Suppose the prevalence of HPV in women 18 - 22 years old is 0.30. What is the probability that in a sample of 60 women from this population 9 or fewer would be infected?

Random variable?

Distribution?

Parameter(s)?

Question?



graph X [weight=PX] if (X<37), hist bin(37) normal gap(3) yscale(0,.12)

Normal Approximation to Binomial

Binomial

- When **np(1-p)** is "large" the normal may be used to approximate the binomial.
- X ~ bin(n,p)

  E(X) = np

  V(X) = np(1-p)
- X is approximately N(np,np(1-p))

Normal Approximation to Binomial

Example

Suppose the prevalence of HPV in women 18 - 22 years old is 0.30. What is the probability that in a sample of 60 women from this population that 9 or less would be infected?

Random variable?

$\Rightarrow$ X = number infected out of 60

Distribution?

$\Rightarrow$ Binomial

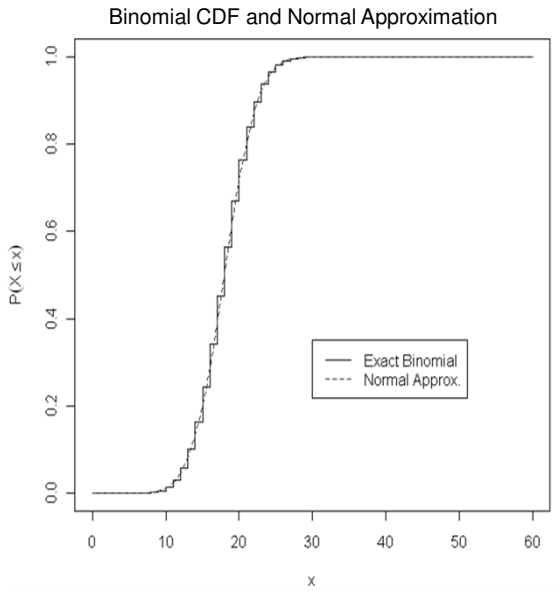
Parameter(s)?

$\Rightarrow$ n = 60, p = .30

Question?

$\Rightarrow$ P(X $\leq$ 9) =

$\Rightarrow$ normal approx. =



P(X $\leq$ 9, n=60, p=.3) = 0.0059

$P\left(\frac{X-np}{\sqrt{np(1-p)}} \leq \frac{9-60*.3}{\sqrt{60*.3*.7}}\right)$=P(Z $\leq$ -2.535) = 0.0056.

Estimation

Estimation

- All probability models depend on parameters. E.g.,
  Binomial depends on probability of success π.
  Normal depends on mean μ, standard deviation σ.
- Parameters are properties of the "population" and are typically unknown.
- The process of taking a sample of data to make inferences about these parameters is referred to as "estimation".
- There are a number of different estimation methods … we will study two estimation methods:

  Maximum likelihood (ML)
  Bayes

Maximum Likelihood

Fisher (1922) invented this general method.

Problem: Unknown model parameters, θ.

Set-up: Write the probability of the data, ***Y***, in terms of the model parameter and the data, $P(Y,\theta)$.

Solution: Choose as your estimate the value of the unknown parameter that makes your data look as likely as possible. Pick $\hat{\theta}$ that maximizes the probability of the observed data.

The estimator $\hat{\theta}$ is called the maximum likelihood estimator (MLE).

Maximum Likelihood - Example

Data: $Y_i = 0/1$ for $i = 1, 2, \ldots n$ (independent)

Model: $Z = \sum_i Y_i$ ~ Binomial(n,π)

Probability: Let's fix the number in the sample at $n = 20$. The resulting model for Z is Binomial with size 20 and success probability π.

The ***probability distribution function*** is:

$$P(Z;\pi) = \binom{20}{Z} \pi^{Z} (1-\pi)^{(20-Z)}$$

where Z is the variable and **π is fixed**.

The ***likelihood function*** is the same function:

$$L(\pi; Z) = \binom{20}{Z} \pi^{Z} (1-\pi)^{(20-Z)}$$

except now π is the variable and **Z is fixed**.

Maximum Likelihood - Example

Two ways to look at this:

- Fix π and look at the probability of different values of Z:

$\pi = 0.1$

Z	$P(Z, \pi)$
0	0.122
1	0.270
2	0.285
3	0.190
4	0.090
5	0.032

- Fix Z and look at the probability under different values of π (this is called the likelihood function):

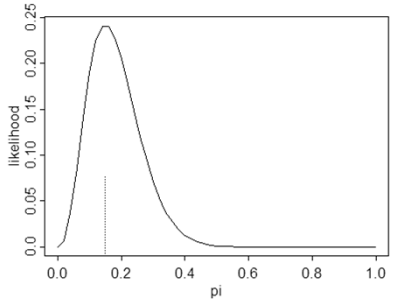
$Z = 3$

π	$P(Z, \pi)$
0.01	0.001
0.05	0.060
0.10	0.190
0.20	0.205
0.30	0.072
0.40	0.012

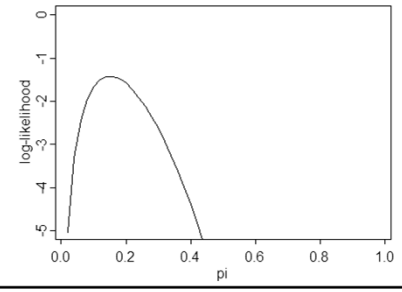
Maximum Likelihood - Example

If you observe the data $Z = 3$ then the likelihood function is shown in the plots below:

P(Z=3) as function of π



log P(Z=3) as function of π



Maximum Likelihood - Example

- We can use elementary calculus (an oxymoron?) to find the maximum of the (log) likelihood function:

$$\frac{d\log L}{d\pi} = 0$$

$$\frac{d}{d\pi} Z\log\pi + (20-Z)\log(1-\pi) = 0$$

$$\frac{Z}{\pi} - \frac{(20-Z)}{1-\pi} = 0$$

$$\hat{\pi} = \frac{Z}{20}$$

- Not surprisingly, the likelihood in this example is maximized at the observed proportion, 3/20.
- Sometimes (e.g. this example) the MLE has a simple closed form. In more complex problems, numerical optimization is used.
- Computers can find these maximum values!

Maximum Likelihood - Notation

$L(\theta)$ = Likelihood as a function of the unknown parameter, θ.

$l(\theta) = \log(L(\theta))$, the log-likelihood.

Usually more convenient to work with analytically and numerically.

$S(\theta) = dl(\theta)/d\theta$ = the "score".

Set $dl(\theta)/d\theta = 0$ and solve for θ to find the MLE.

$I(\theta) = -d^2l(\theta)/d\theta^2$ = the "information".

If evaluated at the MLE, then $-d^2l(\theta)/d\theta^2$ is referred to as the observed information;
$E(-d^2l(\theta)/d\theta^2)$ is referred to as the expected or Fisher information.

$Var(\theta) = I^{-1}(\theta)$ (in most cases)

Maximum Likelihood - Example

$$L(\pi) = \binom{20}{Z} \pi^{Z} (1-\pi)^{(20-Z)}$$

$$\ell(\pi) = Z\log(\pi) + (20-Z)\log(1-\pi)$$

$$S(\pi) = \frac{Z}{\pi} - \frac{(20-Z)}{1-\pi} \quad \Rightarrow \quad \pi = \frac{Z}{20}$$

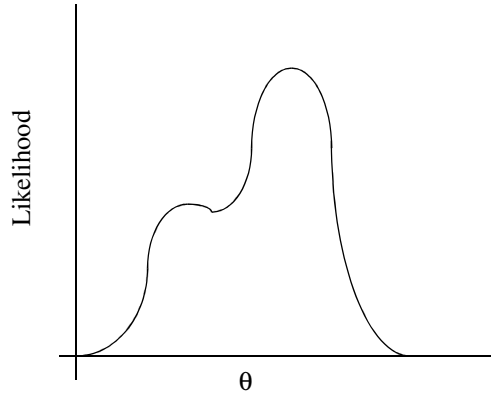
$$I(\pi) = \frac{Z}{\pi^2} + \frac{(20-Z)}{(1-\pi)^2}$$

$$E(I(\pi)) = \frac{20\pi}{\pi^2} + \frac{(20-20\pi)}{(1-\pi)^2} = \frac{20}{\pi(1-\pi)}$$

(note: constant dropped from $\ell(\pi)$)

Numerical Optimization

- In complex problems it may not be possible to find the MLE analytically; in that case we use numerical optimization to search for the value of θ that maximizes the likelihood
- A common problem with maximum likelihood estimation is accidentally finding a local maximum instead of a global one; solution is to try multiple starting values



Comments:

- Maximum likelihood estimates (MLEs) are always based on a probability model for the data.
- Maximum likelihood is the "best" method of estimation for any situation that you are willing to write down a probability model (so generally does not apply to nonparametric problems).
- Maximum likelihood can be used even when there are multiple unknown parameters, in which case θ has several components

  (ie. $\theta_0, \theta_1, \ldots, \theta_p$).
- The MLE is a "point estimate" (i.e. gives the single most likely value of θ). In lecture 5 we will learn about interval estimates, which describe a range of values which are likely to include the true value of θ. We combine the MLE and Var(θ) to generate these intervals.
- The likelihood function lets us compare different models (next).

Model Comparisons

Q: Suppose we have two alternative models for the data; in each case we use maximum likelihood to estimate the parameters. How do we decide which model fits the data "better"?

A: First thought - compare the likelihoods.

- Larger likelihood is better, but …
- the tradeoff is larger likelihood $\Leftrightarrow$ more complex model.
- How to choose?

A common approach is to "penalize" the likelihood for more complex models (i.e. more parameters).

The AIC and BIC are two examples of penalized likelihood measures.

The LOD ("log odds") score can be thought of as a special case (1 parameter) of a penalized likelihood.

Example – LOD scores

Suppose we have a sample of size N gametes in which the number of recombinants (R) and nonrecombinants (N-R) for two loci can be counted. Let θ be the recombination fraction between the two loci. Then the probability of the data can be modeled using the binomial distribution:

$$P(R) = \binom{N}{R} \theta^R (1-\theta)^{N-R}$$

The situation of no linkage corresponds to θ = 0.5, so we can express the models as

Model 1: θ = 0.5

Model 2: θ anywhere between 0 and 0.5

Example – LOD scores

Model 1: The situation of no linkage corresponds to θ = 0.5. If we substitute this into the likelihood equation, we get

$$\begin{aligned} \log_{10} L_1 &= R \log_{10} 0.5 + (N-R) \log_{10} 0.5 \\ &= N \log_{10} 0.5 \end{aligned}$$

This model has 0 (free) parameters.

Model 2: The log-likelihood when θ is unrestricted is

$$\log_{10} L_2 = R \log_{10} \theta + (N-R) \log_{10} (1-\theta)$$

This model has 1 parameter.

Taking the derivative and solving for θ gives

$$\hat{\theta} = \frac{R}{N}$$

If we substitute this back into the log-likelihood, we get …

$$\log_{10} L_2 = R \log_{10} {}^{R}\!/_{N} + (N-R) \log_{10} (1 - {}^{R}\!/_{N})$$

Example – LOD scores

The LOD score is

$$\text{LOD} = (\log_{10} L_2 - \log_{10} L_1)$$

$$= R\log_{10}\left(\frac{R}{N-R}\right) + N\log_{10}\left(\frac{N-R}{0.5N}\right)$$

Large values of the LOD score (> 3) are considered evidence of linkage (i.e. the penalty is 3).

(As we will see, this is a pretty big hurdle to overcome.)

Example – LOD scores

E.g. N = 50 and R = 18

$\hat{\theta}$ = 18/50 = 36%

$\log_{10}L_1$ = -15.0

$\log_{10}L_2$ = -14.2

LOD = -14.2 – (-15.0) = 0.8

$\Rightarrow$ No evidence of linkage; conclude $\theta = .5$

Model Comparisons – AIC, BIC

AIC – Akaike's Information Criterion

BIC – Bayes Information Criterion

AIC = 2 $\ell(\theta)$ - 2k

BIC = 2 $\ell(\theta)$ - klog(n)

k = # parameters

(natural logs now)

- Use to compare a series of models. Pick the model with the largest AIC or BIC
- Larger model ⇒ larger likelihood (typically)
- Therefore, "penalize" the likelihood for each added parameter
- AIC tries to find the model that would have the minimum prediction error on a new set of data.
- BIC tries to find the model with the highest "posterior probability" given the data
- Typically, BIC is more conservative (picks smaller models)

Model Comparisons – AIC, BIC

Example – Recombinants (N=50, R = 18)

log(L1)= -34.66 (natural logs now)

log(L2) = -32.67

	θ = .5	θ arb
AIC	-2*34.66 = **-69.32**	-2*32.67 - 2 = **-67.34**
BIC	-2*34.66 = **-69.32**	-2*32.67 - log(50) = **-69.25**

AIC ⇒ pick θ = .36

BIC ⇒ pick θ = .36 (but almost tied)

Bayes Estimation

Recall Bayes theorem (written in terms of data X and parameter θ):

$$P(\theta|X) = \frac{P(X|\theta)P(\theta)}{\int_{\theta} P(X|\theta)P(\theta)}$$

Notice the change in perspective - θ is now treated as a random variable instead of a fixed number.

P(X|θ) is the likelihood function, as before.

P(θ) is called the *prior distribution* of θ.

P(θ | X) is called the *posterior distribution* of θ.

Based on P(θ | X) we can define a number of possible estimators of θ. A commonly used estimate is the maximum a posteriori (MAP) estimate:

$$\hat{\theta}_{MAP} = \max_{\theta} P(\theta|X)$$

We can also use P(θ | X) to define "credible" intervals for θ.

Bayes Estimation

Comments:

- The MAP estimator is a very simple Bayes estimator. More generally, Bayes estimators minimize a "loss function" – a penalty based on how far $\hat{\theta}$ is from θ (e.g. Loss =$(\hat{\theta} - \theta)^2$).
- The Bayesian procedure provides a convenient way of combining external information or previous data (through the prior distribution) with the current data (through the likelihood) to create a new estimate.
- As N increases, the data (through the likelihood) overwhelms the prior and Bayes estimator typically converges to the MLE
- Controversy arises when P(θ) is used to incorporate subjective beliefs or opinions.
- If the prior distribution P(θ) is simply that θ is uniformly distributed over all possible values, this is called an "uninformative" prior, and the MAP is the same as the MLE.

Bayes Estimation

Example

Suppose a man is known to have transmitted allele A1 to his child at a locus that has only two alleles: A1 and A2. What is his most likely genotype?

Soln. Let X represent the paternal allele in the child and let θ represent the man's genotype:

$X = A1$

$\theta = \{A1A1,\ A1A2,\ A2A2\}$

We can write the likelihood function as:

$P(X \mid \theta = A1A1) = 1$

$P(X \mid \theta = A1A2) = .5$

$P(X \mid \theta = A2A2) = 0$

Therefore, the MLE is $\theta = A1A1$.

Bayes Estimation

Suppose, however, that we know that the frequency of the A1 allele in the general population is only 1%. Assuming HW equilibrium we have

$P(\theta = A1A1) = .0001$

$P(\theta = A1A2) = .0198$

$P(\theta = A2A2) = .9801$

This leads to the posterior distribution

$$P(\theta = A1A1 \mid X) = P(X \mid \theta = A1A1)\, P(\theta = A1A1) / P(X) = 1 * .0001 / .01 = .01$$

$$P(\theta = A1A2 \mid X) = P(X \mid \theta = A1A2)\, P(\theta = A1A2) / P(X) = .5 * .0198 / .01 = .99$$

$P(\theta = A2A2 \mid X) = 0$

So the Bayesian MAP estimator is $\theta = A1A2$.

Exercise: redo assuming the man has 2 children who both have the A1 paternal allele.

Summary

- Maximum likelihood is a method of estimating parameters from data
- ML requires you to write a probability model for the data
- MLE's may be found analytically or numerically
- (Inverse of the negative of the) second derivative of the log-likelihood gives variance of estimates
- Comparison of log-likelihoods allows us to choose between alternative models
- Bayesian procedures allow us to incorporate additional information about the parameters in the form of prior data, external information or personal beliefs.

Problem 1

Suppose we are interested in estimating the recombination fraction, θ, from the following experiment. We do a series of crosses: AB/ab x AB/ab and measure the frequency of the various phases in the gametes (assume we can do this). If the recombination fraction is θ then we expect the following probabilities (sorry, I can't explain these…):

phase	probability (*4)
AB	$3 - 2\theta + \theta^2$
Ab	$2\theta - \theta^2$
aB	$2\theta - \theta^2$
ab	$1 - 2\theta + \theta^2$

Suppose we observe (AB,Ab,aB,aa) = (125,18,20,34). Use maximum likelihood to estimate θ.

Solution to problem 1

$\Pr(\text{data} \mid \theta) \propto (3\text{-}2\theta+\theta^2)^{AB}\,(2\theta - \theta^2)^{Ab}\,(2\theta - \theta^2)^{aB}\,(1\text{-}2\theta+\theta^2)^{ab}$

$l(\theta) = AB \log(3\text{-}2\theta+\theta^2) + (Ab+aB) \log(2\theta - \theta^2) + ab \log(1\text{-}2\theta+\theta^2)$

$$\frac{d\ell(\theta)}{d\theta} = \frac{2AB(\theta-1)}{3-2\theta+\theta^2} + \frac{2(Ab+aB)(1-\theta)}{2\theta-\theta^2} + \frac{2ab(\theta-1)}{1-2\theta+\theta^2} = 0$$

Numerical solution gives $\theta = .21$

$$\frac{d^2\ell(\theta)}{d\theta^2} = \frac{AB(1+2\theta-\theta^2)}{[3-2\theta+\theta^2]^2} - \frac{(Ab+aB)}{\theta^2} - \frac{ab}{(1-\theta)^2}$$

$$I = E\left(-\frac{d^2\ell(\theta)}{d\theta^2}\right) = -\mathrm{N} * \left(\frac{1+2\theta-\theta^2}{3-2\theta+\theta^2} + \frac{4(1-\theta)}{\theta} + 1\right)$$

$= \mathrm{N}*16.6$

$\mathrm{Var}(\theta) = 1/213.6 = .00468$

Problem 2

Every human being can be classified into one of four blood groups: O, A, B, AB. Inheritance of these blood groups is controlled by 1 gene with 3 alleles: O, A and B where O is recessive to A and B. Suppose the frequency of these alleles is r, p, and q, respectively (p+q+r=1). If we observe (O,A,B,AB) = (176,182,60,17) use maximum likelihood to estimate r, p and q.

Solution to problem 2

First, we use basic genetics to find the probability of the observed phenotypes in terms of the unknown parameters. Assuming random mating, we have:

Genotype	prob.	Phenotype	prob.
OO	r^2	O	r^2
AA	p^2		
AO	2pr	A	$p^2 + 2pr$
BB	q^2		
BO	2qr	B	$q^2 + 2qr$
AB	2pq	AB	2pq

$\Pr(\text{data} \mid \theta) \propto (r^2)^O (p^2+2pr)^A (q^2+2qr)^B (2pq)^{AB}$

$l(p,q,r) = 2O\log(r) + A\log(p^2+2pr) + B\log(q^2+2qr) + AB\log(p) + AB\log(q)$

To estimate p, q and r, we need to maximize $l(p,q,r)$ subject to the constraint p+q+r=1. This constraint makes the problem a bit harder …. one approach is to just put r = 1-p-q in the likelihood so we have just 2 parameters … p and q. Then

$$\frac{dl}{dp} = -\frac{2O}{r} + \frac{2Ar}{p(2r+p)} - \frac{2Bq}{q(2r+q)} + \frac{AB}{p} = 0$$

$$\frac{dl}{dq} = -\frac{2O}{r} - \frac{2Ap}{p(2r+p)} + \frac{2Br}{q(2r+q)} + \frac{AB}{q} = 0$$

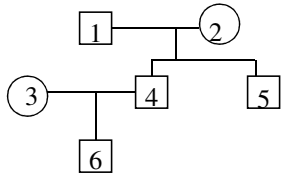
For (O,A,B,AB) = (176,182,60,17), this gives

p = .264 q = .093 r = .642

Further analysis would take 2nd derivatives to find the information and, therefore, the variances of the estimates.

Problem 3

Suppose we have the following simple pedigree.



Define the phenotype of person i as H_i and the genotype as G_{iH} How can we use maximum likelihood to estimate parameters of the *penetrance function,* $\Pr(H \mid G; \theta)$?

Solution to problem 3

- If we knew all the genotypes the problem would be "easy". We would simply write down the log-likelihood and maximize it numerically or analytically:

$$l(\theta) = \sum_i \log \Pr(H_i \mid G_i)$$

- If we don't know the genotypes (only data are the phenotypes), then we must maximize

$$l(\theta) = \log \Pr(H)$$

where H represents the collection of all 6 phenotypes. The general idea is to use the total probability rule to write

$$\Pr(H) = \sum_G \Pr(H \mid G)\Pr(G)$$
$$= \sum_{G_1,G_2,G_3,G_4,G_5,G_6} \left\{ \prod_i \Pr(H_i \mid G_i) \right\} \Pr(G_1, G_2, G_3, G_4, G_5, G_6)$$

Further simplification is achieved by writing

$$\Pr(G_1,G_2,G_3,G_4,G_5,G_6) = \Pr(G_6 \mid G_3,G_2,G_3,G_4,G_5)\Pr(G_5 \mid G_1,G_2,G_3,G_4)\Pr(G_4 \mid G_1,G_2,G_3,G_4)\times \Pr(G_3 \mid G_1,G_2)\Pr(G_2 \mid G_1)\Pr(G_1)$$

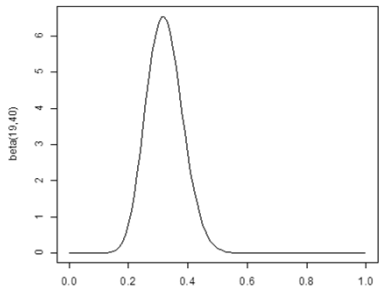
Since the genotype of each individual is determined only by his/her parents

$$\Pr(G_1,G_2,G_3,G_4,G_5,G_6) = \Pr(G_6 \mid G_3,G_4)\Pr(G_5 \mid G_1,G_2)\Pr(G_4 \mid G_1,G_2)\Pr(G_3)\Pr(G_2)\Pr(G_1)$$

Given the inheritance probabilities ($\Pr(G_i \mid G_j, G_k)$) and population frequencies of the genotypes ($\Pr(G_i)$), we have a fully specified model and can maximize the likelihood using a computer.

Problem 4

Suppose we wish to estimate the recombination fraction for a particular locus. We observe N = 50 and R = 18. Several previously published studies of the recombination fraction in nearby loci (that we believe should have similar recombination fractions) have shown recombination fractions between .22 and .44. We decide to model this prior information as a beta distribution (see http://en.wikipedia.org/wiki/Beta_distribution) with parameters a = 19 and b = 40:



Find the MLE and Bayesian MAP estimators of the recombination fraction. Also find a 95% confidence interval (for the MLE) and a 95% credible interval (for the MAP)

Solution to problem 4

The data follow a binomial distribution with N = 50, R = 18 and the prior information is captured by a beta distribution with parameters a = 19, b = 40:

$$P(\theta) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1}$$

$$P(X \mid \theta) = \frac{N!}{R!(N-R)!}\theta^{R}(1-\theta)^{N-R}$$

Working through Bayes theorem, we find …

$$P(\theta \mid X) = \frac{\Gamma(N+a+b)}{\Gamma(a+R)\Gamma(N-R+b)}\theta^{a+R-1}(1-\theta)^{N-R+b-1}$$

which is another beta distribution with parameters (a+R) and (N-R+b). The mode of the beta distribution with parameters α and β is $(\alpha-1)/(\alpha+\beta-2)$ so

$$\hat{\theta}_{MAP} = \frac{a+R-1}{N+a+b-2} = \frac{36}{107} = .336$$

Also, we can find the 2.5th and 97.5th percentiles of the posterior distribution (95% credible interval): [.23 - .40]

For comparison the MLE is 18/50 = 0.36 with a 95% confidence interval of [.23 - .49]

Sampling Distributions

Sample Summaries

Population

- Size N (usually ∞)
- Mean = μ

$$\mu = \sum p_j X_j \quad or \quad \int \ldots$$

- Variance = σ^2

$$\sigma^2 = \sum p_j \left(X_j - \mu\right)^2 \quad or \quad \int \ldots$$

Sample

- Size n
- Mean = $\overline{X}$

$$\overline{X} = \frac{1}{n} \sum_{j=1}^{n} X_j$$

- Sample variance = s^2

$$s^2 = \frac{1}{n-1} \sum_{j=1}^{n} \left(X_j - \overline{X}\right)^2$$

Sums of Normal Random Variables

We already know that linear functions of a normal rv are normal. What about combinations (eg. sums) of normals?

==> If $X_j \sim N(\mu_j, \sigma_j^2)$ (indep) then

$$Y = \sum_{j=1}^{n} X_j$$

$$Y \sim N\left(\sum_{j=1}^{n} \mu_j, \sum_{j=1}^{n} \sigma_j^2\right)$$

Combine this with what we have learned about linear functions of means and variances to get ...

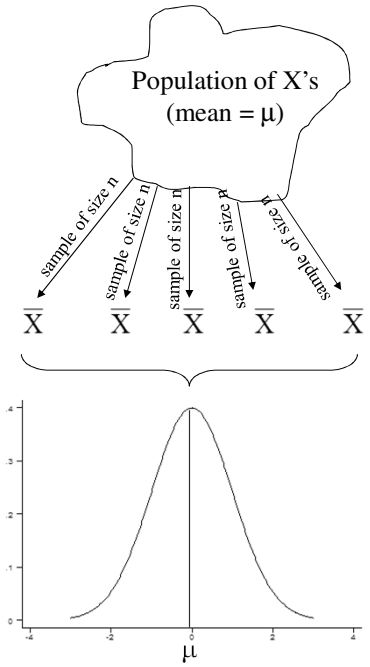
Distribution of the Sample Mean

A. When sampling from a **normally** distributed population:

1. The distribution of $\overline{X}$ is **normal**.
2. $\overline{X}$ is a random variable.
3. Mean of $\overline{X}$ is $\mu_{\overline{X}}$ which equals μ, the mean of the population.
4. Variance of $\overline{X}$ is $\sigma^2_{\overline{X}}$ which equals, σ^2/n the variance of the population divided by the sample size.
5. $\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$.

B. When the population is **non-normal** but the sample size is large, the **Central Limit Theorem** applies.

Distribution of the Sample Mean



Central Limit Theorem

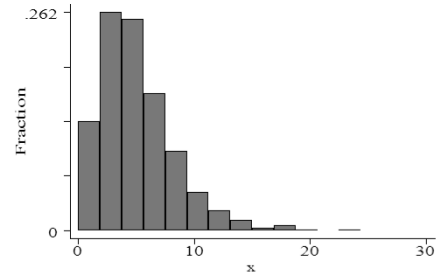
Given a population with any non-normally distributed variables with a mean μ and a variance σ^2, then for large enough sample sizes, the distribution of the sample mean, $\overline{X}$, will be **approximately normal** with means μ and variance σ^2/n.

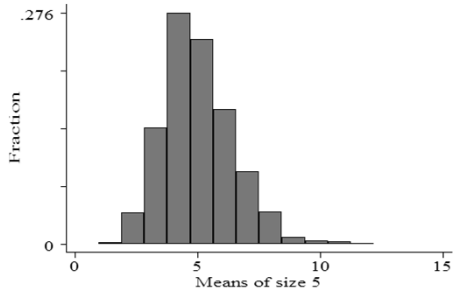
$$\text{n large} \rightarrow \overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

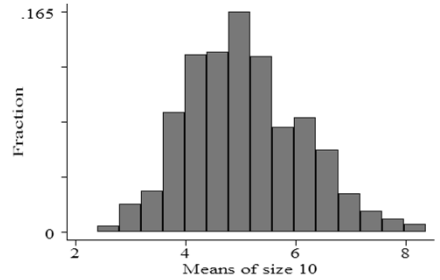
- In general, this applies for $n \geq 30$.
- As n increases, the normal approximation improves.

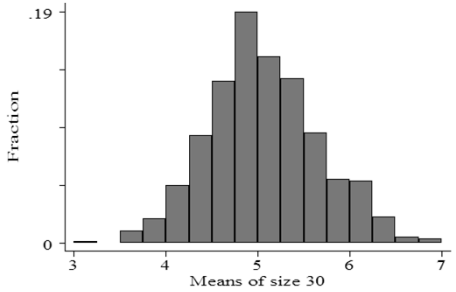
Central Limit Theorem - Illustration

Population









Distribution of Sample Mean

In applications we can address:

What is the probability of obtaining a sample with mean larger (smaller) than T (some constant) when sampling from a population with mean μ and variance σ^2?

Transform to Standard Normal

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

random variable	=	$\overline{X}$
distribution of sample mean	≈	Normal
expected value of sample mean	=	μ
standard deviation of sample mean	=	$\frac{\sigma}{\sqrt{n}}$

Distribution of the Sample Mean

EXAMPLE:

Suppose that for Seattle sixth grade students the mean number of missed school days is 5.4 days with a standard deviation of 2.8 days. What is the probability that a random sample of size 49 (say Ridgecrest's 6th graders) will have a mean number of missed days greater than 6 days?

Random Variable

Distribution

Parameters

Question

Find the probability that a random sample of size 49 from this population will have a mean greater than 6 days.

μ = 5.4 days

σ = 2.8 days

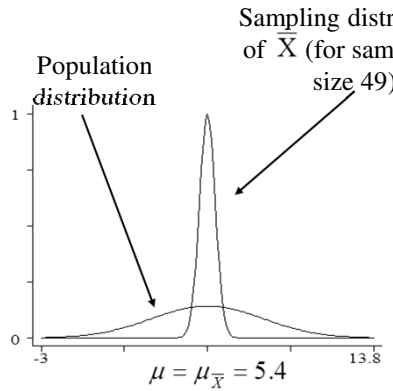
n = 49

$$\sigma_{\overline{X}} = \sigma / \sqrt{n} = 2.8 / \sqrt{49} = 0.4$$

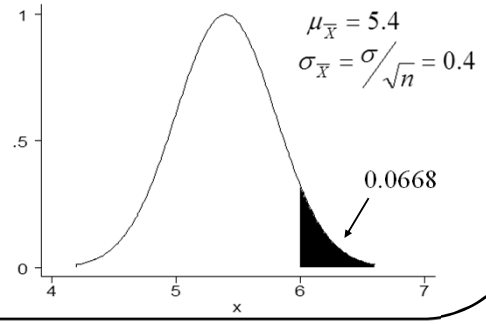
$$\mu_{\overline{X}} = 5.4$$

$$P(\overline{X} > 6) = P\left(\frac{\overline{X} - \mu_{\overline{X}}}{\sigma_{\overline{X}}} > \frac{6 - 5.4}{0.4}\right)$$

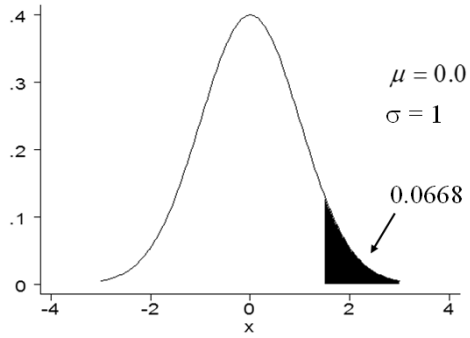
$$= P(Z > 1.5) = 0.0668$$



Let's look at the sampling distribution more closely ...



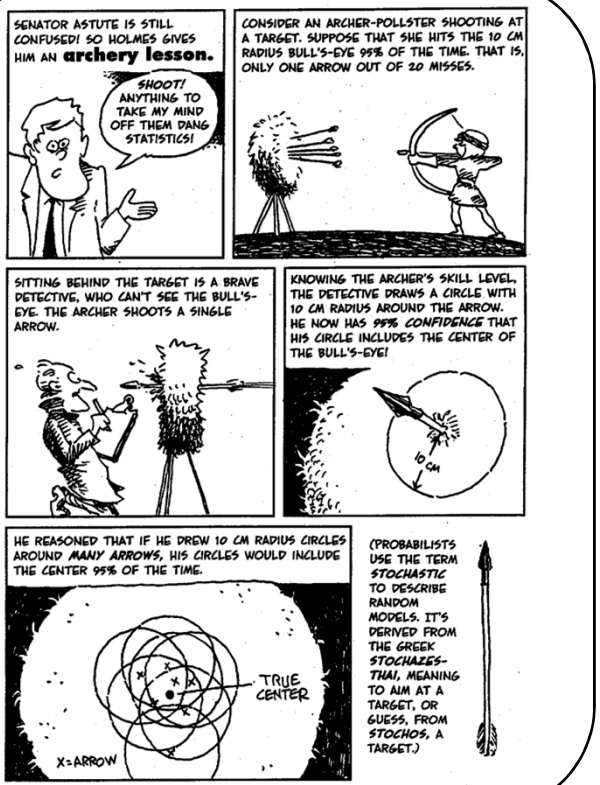
In terms of the standard normal ...



What is the probability that a random sample (size 49) from this population has a mean between 4 and 6 days? Check that ….

$$\begin{aligned} P(4 \le \overline{X} \le 6) &= P(-3.5 \le Z \le 1.5) \\ &= P(Z \le 1.5) - P(Z \le -3.5) \\ &= .933 \end{aligned}$$

Confidence Intervals



Confidence Intervals

Q: When we do not know the population parameter, how can we use the sample to estimate the population mean, and use our knowledge of probability to give a range of values consistent with the data?

Parameter: μ

Estimate: $\overline{X}$

Given a normal population, or large sample size, we can state:

$$P\left[-1.96 \le \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \le +1.96\right] = 0.95$$

Confidence Intervals

$$P\left[-1.96 \le \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \le +1.96\right] = 0.95$$

We can do some rearranging:

$$P\left[-1.96\sigma/\sqrt{n} \le \overline{X}-\mu \le +1.96\sigma/\sqrt{n}\right] = 0.95$$

$$P\left[-\overline{X}-1.96\sigma/\sqrt{n} \le -\mu \le -\overline{X}+1.96\sigma/\sqrt{n}\right] = 0.95$$

$$P\left[\overline{X}-1.96\sigma/\sqrt{n} \le \mu \le \overline{X}+1.96\sigma/\sqrt{n}\right] = 0.95$$

The interval

$$\left(\overline{X}-1.96\sigma/\sqrt{n}\,,\ \overline{X}+1.96\sigma/\sqrt{n}\right)$$

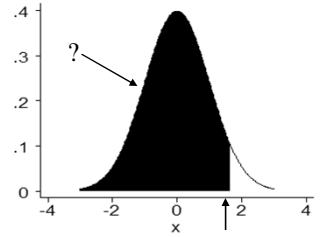
is called a **95% confidence interval for** μ.

Normal Quantiles

Go back to Rosner, table 3 ….

Notice that we can use the table two ways:

(1) Given a particular x value (the quantile) we can look up the probability:

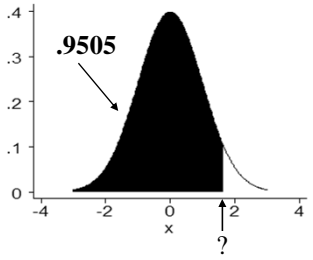


P(Z ≤ **1.65**) = ?

Rosner table 3 …

x	A	B	C	D
:	:	:	:	:
1.65	.9505	.0495	.4505	.9011
:	:	:	:	:

(2) Given a particular probability, we can look up the quantile:



$P(Z \leq ?) = \mathbf{.9505}$

x	A	B	C	D
1.65	**.9505**	.0495	.4505	.9011

$Q_Z^{(p)}$ is the value of x such that $P(Z \leq x) = p$

Verify: $Q_Z^{(.95)} = 1.65$

$Q_Z^{(.975)} = 1.96$

$P(Q_Z^{(.05)} \leq Z \leq Q_Z^{(.95)}) = .90 \quad \Rightarrow \quad Q_Z^{(.05)} = -1.65,\ Q_Z^{(.95)} = 1.65$

Notice that $Q_Z^{(p)} = -Q_Z^{(1-p)}$

Confidence Intervals

σ known

When σ is known we can construct a confidence interval for the population mean, μ, for any given confidence level, (1 - α). Instead of using 1.96 (as with 95% CI's) we simply use a different constant that yields the right probability.

So if we desire a (1 - α) confidence interval we can derive it based on the statement

$$P\left[Q_Z^{\left(\frac{\alpha}{2}\right)} < \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} < Q_Z^{\left(1-\frac{\alpha}{2}\right)}\right] = 1 - \alpha$$

That is, we find constants $Q_Z^{\left(\frac{\alpha}{2}\right)}$ and $Q_Z^{\left(1-\frac{\alpha}{2}\right)}$ that have exactly (1 - α) probability between them.

A (1 - α) Confidence Interval for the Population Mean

$$\left(\overline{X} + Q_Z^{\left(\frac{\alpha}{2}\right)} \times \frac{\sigma}{\sqrt{n}}, \overline{X} + Q_Z^{\left(1-\frac{\alpha}{2}\right)} \times \frac{\sigma}{\sqrt{n}}\right)$$

Confidence Intervals
σ known - EXAMPLE

Suppose gestational times are normally distributed with a standard deviation of 6 days. A sample of 30 second time mothers yield a mean pregnancy length of 279.5 days. Construct a 90% confidence interval for the mean length of second pregnancies based on this sample.

Confidence Intervals
σ unknown

To get a CI for μ using the methods outlined above, we need $\overline{X}$ and σ^2. But usually, **σ is unknown** - we only have $\overline{X}$ and s^2. It turns out that even though

$$\frac{(\overline{X}-\mu)}{\sigma/\sqrt{n}}$$

is normally distributed,

$$\frac{(\overline{X}-\mu)}{s/\sqrt{n}}$$

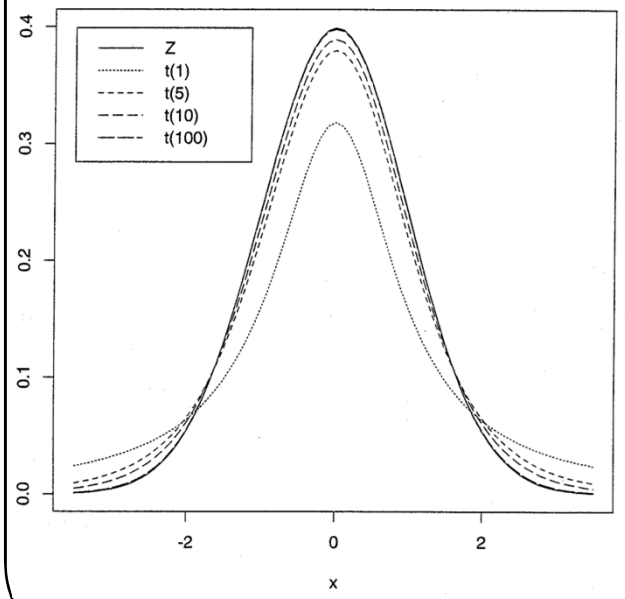
is not (quite)!

W.S. Gosset worked for Guinness Brewing in Dublin, IR. He was forced to publish under the pseudonym "Student". In 1908 he derived the distribution of

$$\frac{(\overline{X}-\mu)}{s/\sqrt{n}}$$

which is now known as Student's **t-distribution**.

Normal and t distributions



Confidence Intervals
σ^2 unknown
t Distribution

When σ is unknown we replace it with the estimate, s, and use the t-distribution. The statistic

$$\frac{\overline{X}-\mu}{s/\sqrt{n}}$$

has a t-distribution with n-1 *degrees of freedom*.

We can use this distribution to obtain a confidence interval for μ even when σ is not known.

See Rosner, table 5 or `display tprob(df,t)`

A (1-α) **Confidence Interval for the Population Mean when σ is unknown**

$$\left(\overline{X}+Q_{t(n-1)}^{\left(\frac{\alpha}{2}\right)}\times s/\sqrt{n},\ \overline{X}+Q_{t(n-1)}^{\left(1-\frac{\alpha}{2}\right)}\times s/\sqrt{n}\right)$$

Confidence Intervals - σ^2 unknown
t Distribution - EXAMPLE

Given our 30 moms with a mean gestation of 279.5 days and a variance of 28.3 days^2, we can now compute a 95% confidence interval for the mean length of pregnancies for second time mothers:

Confidence Intervals - sample variance

Q: Can we derive a confidence interval for the sample variance?

A: Yes. We'll need the **Chi-square distribution**

Definition: The sum of squared independent standard normal random is a random variable with a **Chi-square** distribution with n degrees of freedom.

Let Z_i be standard normals, N(0,1). Let

$$X = Z_1^2 + Z_2^2 + \ldots + Z_n^2 = \sum_{i=1}^{n} Z_i^2$$

X has a χ^2(n) distribution

Chi-square Distribution

Properties of χ^2 (n): Let X ~ χ^2(n).

1. X ≥ 0
2. E[X] = n
3. V[X] = 2n
4. **n**, the parameter of the distribution is called *the degrees of freedom*.

Chi-square Distribution
Sample Variance

The Chi-square distribution describes the distribution of the **sample variance**. Recall

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$$

and

$$(n-1)\frac{s^2}{\sigma^2} = \sum_{i=1}^{n}\left(\frac{X_i - \overline{X}}{\sigma}\right)^2$$

Now the right side almost looks like

$$\sum_{i=1}^{n}\left(\frac{X_i - \mu}{\sigma}\right)^2$$

which would be χ^2(n).

Since μ is estimated by $\overline{X}$ one degree of freedom is lost leading to …

$$(n-1)\frac{s^2}{\sigma^2} \sim \chi^2$$ with n-1 degrees of freedom

Chi-square Distribution
Confidence Interval for σ^2

We can use the Chi-square distribution to obtain a (1 - α) confidence interval for the **population variance**.

$$P\left[Q_{\chi^2_{(n-1)}}^{\left(\frac{\alpha}{2}\right)} < (n-1)\frac{s^2}{\sigma^2} < Q_{\chi^2_{(n-1)}}^{\left(1-\frac{\alpha}{2}\right)}\right] = 1-\alpha$$

Now, inverting this statement yields:

$$P\left[s^2 \times (n-1)/Q_{\chi^2_{(n-1)}}^{\left(1-\frac{\alpha}{2}\right)} < \sigma^2 < s^2 \times (n-1)/Q_{\chi^2_{(n-1)}}^{\left(\frac{\alpha}{2}\right)}\right] = 1-\alpha$$

Therefore,

A (1 - α) **Confidence Interval for the Population Variance**

$$\left(s^2 \times (n-1)/Q_{\chi^2_{(n-1)}}^{\left(1-\frac{\alpha}{2}\right)},\ s^2 \times (n-1)/Q_{\chi^2_{(n-1)}}^{\frac{\alpha}{2}}\right)$$

Chi-square Distribution
Confidence Interval for σ^2 - EXAMPLE

Suppose for the second time mothers were not happy using the standard deviation of 6 days since it was based on the population of all mothers regardless of parity. The sample variance was 28.3 days^2. What is a 95% confidence interval for the variance of the length of second pregnancies?

Summary

- General (1 - α) Confidence Intervals.
- CI for μ, σ assumed known $\rightarrow$ Z.
- CI for μ, σ unknown $\rightarrow$ T.
- CI for $\sigma^2 \rightarrow \chi^2$

- $\uparrow$confidence $\rightarrow$ wider interval
- $\uparrow$sample size $\rightarrow$ narrower interval

Hypothesis Testing

The ideas in hypothesis testing are based on **deductive reasoning** - we assume that some probability model is true and then ask "What are the chances that these observations came from that probability model?".

Hypothesis Testing

- Null Hypothesis
- Alternative Hypothesis
- Significance level
- Statistically significant
- Critical value
- Acceptance / rejection region
- p-value
- power
- Types of errors: Type I (α), Type II (β)
- One-sided (one-tailed) test
- Two-sided (two-tailed) test

Hypothesis Testing
Motivation

1. Is the chance of getting a cold different when subjects take vitamin C than when they take placebo? (Pauling 1971 data).

2. Suppose that 6 out of 15 students in a grade-school class develop influence, whereas 20% of grade-school children nationwide develop influenza. Is there evidence of an excessive number of cases in the class?

Hypothesis Testing
Motivation

3. In a study of 25 hypertensive men we find a mean serum-cholesterol level of 220 mg/ml. In the 20-74 year-old male population the mean serum cholesterol is 211 mg/ml with standard deviation of 46 mg/ml.

- Is the mean for the population of hypertensive men also 211 mg/ml?
- Is the data consistent with that model?
- What if $\overline{X}$ = 230 mg/ml?
- What if $\overline{X}$ = 250 mg/ml?
- What if the sample was of 100 instead of 25?

Hypothesis Testing

Define:

μ = population mean serum cholesterol for male hypertensives

Hypothesis:

1. Null Hypothesis: Generally, the hypothesis that the unknown parameter equals a fixed value.

H_0: $\mu = 211$ mg/ml

2. Alternative Hypothesis: contradicts the null hypothesis.

H_A: $\mu \neq 211$ mg/ml

Hypothesis Testing

Decision / Action:

We assume that either H_0 or H_A is true. Based on the data we will choose one of these hypotheses.

	H_0 Correct	H_A Correct
Decide H_0	$1-\alpha$	β
Decide H_A	α	$1-\beta$

α = **significance level**
$1 - \beta$ = **power**

Hypothesis Testing

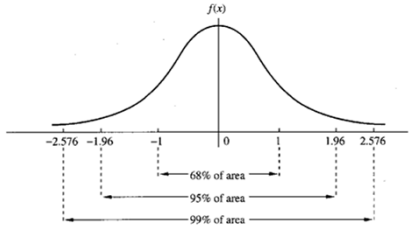
Let's fix α, for example, $\alpha = 0.05$.

$0.05 = \alpha = P[\text{ choose } H_A \mid H_0 \text{ true }]$

$\alpha = P[\text{ reject } H_0 \mid H_0 \text{ true }]$

Q: How to construct a procedure that makes this error with only 0.05 probability?

A: Suppose we assume H_0 is true and suppose that, using that assumption, the data should give us a standard normal, Z.



If $\mu = 0$ then $|Z|$ is rarely "large". A "large" $|Z|$ would make me question whether $\mu = 0$.

Hypothesis Testing

Therefore, **we reject** H_0 **if** $|Z| > 1.96$.

$\alpha = P[\text{reject } H_0 \mid H_0 \text{ true}] = 0.05$

Then if we do find a large value of $|Z|$ we can claim that:

- **Either** H_0 is true and something unusual happened (with probability α)…
- **or**, H_0 is not true.

Given α and H_0 we can construct a test of H_0 with a specified significance level. But remember, we start by assuming that H_0 is true - we haven't proved it is true. Therefore, we usually say

- $|Z| > 1.96$ then we **reject** H_0.
- $|Z| < 1.96$ then we **fail to reject** H_0.

Hypothesis Testing

Cholesterol Example:

Let μ be the mean serum cholesterol level for male hypertensives. We observe

$\overline{X}$ = 220 mg/ml

Also, we are told that for the general population...

μ_0 = mean serum cholesterol level for males = 211 mg/ml

σ = std. dev. of serum cholesterol for males = 46 mg/ml

NULL HYPOTHESIS: mean for male hypertensives is the same as the general male population.

ALTERNATIVE HYPOTHESIS: mean for male hypertensives is different than the mean for the general male population.

$H_0 : \mu = \mu_0 = 211$ mg/ml

$H_A : \mu \neq \mu_0$ ($\mu \neq 211$ mg/ml)

Hypothesis Testing

Cholesterol Example:

Test H_0 with significance level α.

Under H_0 we know:

$$\frac{\overline{X} - \mu_o}{\sigma / \sqrt{n}} \sim N(0,1)$$

Therefore,

•**Reject** H_0 if $\left|\frac{\overline{X} - \mu_0}{\sigma / \sqrt{n}}\right| > 1.96$ gives an $\alpha = 0.05$ test.

•**Reject** H_0 if

$$\overline{X} > \mu_0 + 1.96\frac{\sigma}{\sqrt{n}} \quad \text{or}$$

$$\overline{X} < \mu_0 - 1.96\frac{\sigma}{\sqrt{n}}$$

Hypothesis Testing

Cholesterol Example:

TEST: **Reject** H_0 if

$$\overline{X} > 211 + 1.96\frac{46}{\sqrt{25}} \quad \text{or}$$

$$\overline{X} < 211 - 1.96\frac{46}{\sqrt{25}}$$

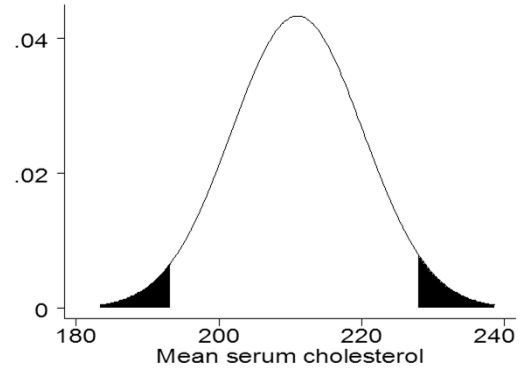
$$\overline{X} > 228.03 \quad \text{or}$$

$$\overline{X} < 192.97$$

In terms of Z ...

$$Z = \frac{\overline{X} - \mu_0}{\sigma / \sqrt{n}}$$

Reject H_0 if Z<-1.96 or Z> 1.96



Hypothesis Testing

p-value:

- smallest possible α for which the observed sample would still reject H_0.
- probability of obtaining a result as extreme or more extreme than the actual sample (give H_0 true).

NOTE: probability calculations are always based on a model.

Hypothesis Testing

p-value: Cholesterol Example

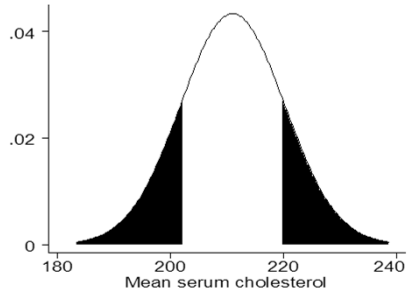
$\overline{X}$ = 220 mg/ml n = 25 σ = 46 mg/ml

$H_0 : \mu = 211$ mg/ml

$H_A : \mu \neq 211$ mg/ml

p-value is given by:

$2 * P[\overline{X} > 220] = .33$



Determination of Statistical Significance for Results from Hypothesis Tests

Either of the following methods can be used to establish whether results from hypothesis tests are statistically significant:

(1) The test statistic Z can be computed and compared with the critical value $Q_z^{(1-\alpha/2)}$ at an α level of .05. Specifically, if H_0: $\mu = \mu_0$ versus H_1: $\mu \neq \mu_0$ are being tested and $|Z| > 1.96$, then H_0 is rejected and the results are declared *statistically significant* (i.e., $p < .05$). Otherwise, H_0 is accepted and the results are declared *not statistically significant* (i.e., $p \geq .05$). We refer to this approach as the **critical-value method**.

(2) The exact p-value can be computed, and if $p < .05$, then H_0 is rejected and the results are declared *statistically significant* . Otherwise, if $p \geq .05$ then H_0 is accepted and the results are declared *not statistically significant* . We will refer to this approach as the **p-value method** .

Guidelines for Judging the Significance of p-value (Rosner pg 200)

If $.01 \leq p < .05$, then the results are *significant*.

If $.001 \leq p < .01$, then the results are *highly significant*.

If $p < .001$, then the results are *very highly significant*.

If $p > .05$, then the results are considered *not statistically significant* (sometimes denoted by NS).

However, if $.05 \leq p < .10$, then a trend toward statistically significance is sometimes noted.

Hypothesis Testing and Confidence Intervals

Hypothesis Test: Fail to reject H_0 if

$$\overline{X} < \mu_0 + Q_Z^{1-\frac{\alpha}{2}} \times \frac{\sigma}{\sqrt{n}}$$

and $$\overline{X} > \mu_0 - Q_Z^{1-\frac{\alpha}{2}} \times \frac{\sigma}{\sqrt{n}}$$

Confidence Interval: Plausible values for μ are given by

$$\mu < \overline{X} + Q_Z^{1-\frac{\alpha}{2}} \times \frac{\sigma}{\sqrt{n}}$$

and $$\mu > \overline{X} - Q_Z^{1-\frac{\alpha}{2}} \times \frac{\sigma}{\sqrt{n}}$$

Hypothesis Testing
"how many sides?"

Depending on the alternative hypothesis a test may have a **one-sided alternative** or a **two-sided alternative**. Consider

$$H_0 : \mu = \mu_0$$

We can envision (at least) three possible alternatives

$$H_A : \mu \neq \mu_0 \qquad (1)$$

$$H_A : \mu < \mu_0 \qquad (2)$$

$$H_A : \mu > \mu_0 \qquad (3)$$

(1) is an example of a "two-sided alternative"

(2) and (3) are examples of "one-sided alternatives"

The distinction impacts

- Rejection regions
- p-value calculation

Hypothesis Testing
"how many sides?"

Cholesterol Example: Instead of the two-sided alternative considered earlier we may have only been interested in the alternative that hypertensives had a higher serum cholesterol.

$$H_0 : \mu = 211$$

$$H_A : \mu > 211$$

Given this, an $\alpha = 0.05$ test would reject when

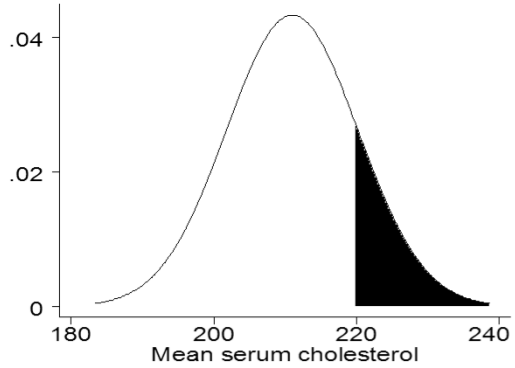
$$\frac{\overline{X} - \mu_0}{\sigma / \sqrt{n}} = Z > Q_Z^{(1-0.05)} = 1.65$$

We put all the probability on "one-side".

The p-value would be half of the previous,

$$\text{p-value} = P[\overline{X} > 220]$$

$$= .163$$



Hypothesis Testing

Through this worked example we have seen the basic components to the statistical test of a scientific hypothesis.

Summary

1. Identify H_0 and H_A
2. Identify a test statistic
3. Determine a significance level, $\alpha = 0.05$, $\alpha = 0.01$
4. Critical value determines rejection / acceptance region
5. p-value
6. Interpret the result

Contingency Tables

Overview

1) **Types of Variables**
2) **Comparing (2) Categorical Variables**
   - Contingency (two-way) tables
   - χ^2 Tests
3) **2 x 2 Tables**
   - Sampling designs
   - Testing for association
   - Estimation of effects
   - Paired binary data
4) **Stratified Tables**
   - Confounding
   - Effect Modification

Factors and Contingency Tables

Definition: A **factor** is a categorical (discrete) variable taking a small number of values that represent the levels of the factor.

Examples

Gender with two levels: 1 = Male and 2 = Female

Disease status with three levels: 1 = Progression, 2 = Stable, 3 = Improved

AgeFactor with 4 levels: 1 = 20-29 yrs, 2 = 30-39, 3 = 40-49, 4 = 50-59

Factors and Contingency Tables

Data description: Form one-way, two-way or multi-way tables of frequencies of factor levels and their combinations

- To assess whether two factors are related, we often construct an R x C table that cross-classifies the observations according to the 2 factors.
- Examining two-way tables of Factor A vs Factor B at each level of a third Factor C shows how the A/B association may be explained or modified by C (later).

Data Summary: Categorical data are often summarized by reporting the proportion or percent in each category. Alternatively, one sometimes sees a summary of the relative proportion (odds) in each category (relative to a "baseline" category).

Testing: We can test whether the factors are related using a χ^2 test.

Categorical Data

Example: From Doll and Hill (1952) - retrospective assessment of smoking frequency. The table displays the daily average number of cigarettes for lung cancer patients and control patients. Note there are equal numbers of cancer patients and controls.

	Daily # cigarettes						
	None	< 5	5-14	15-24	25-49	50+	Total
Cancer	7	55	489	475	293	38	1357
	0.5%	4.1%	36.0%	35.0%	21.6%	2.8%	
Control	61	129	570	431	154	12	1357
	4.5%	9.5%	42.0%	31.8%	11.3%	0.9%	
Total	68	184	1059	906	447	50	2714

χ^2 Test

We want to test whether the smoking frequency is the same for each of the populations sampled. We want to test whether the **groups** are **homogeneous** with respect to a characteristic.

H_0: smoking probability same in both groups

H_A: smoking probability not the same

Q: What does H_0 predict we would observe if all we knew were the marginal totals?

	Daily # cigarettes						
	None	< 5	5-14	15-24	25-49	50+	Total
Cancer							1357
Control							1357
Total	68	184	1059	906	447	50	2714

χ^2 Test

A: H_0 predicts the following **expectations**:

	Daily # cigarettes						
	None	< 5	5-14	15-24	25-49	50+	Total
Cancer	34	92	529.5	453	223.5	25	1357
Control	34	92	529.5	453	223.5	25	1357
Total	68	184	1059	906	447	50	2714

Each group has the same proportion in each cell as the overall **marginal proportion.** The "equal" expected number for each group is the result of the equal sample size in each group (what would change if there were half as many cases as controls?)

χ^2 Test

Summing the differences between the observed and expected counts provides an overall assessment of H_0.

$$X^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \sim \chi^2((r-1)\times(c-1))$$

X^2 is known as the **Pearson's Chi-square Statistic.**

- Large values of X^2 suggests the data are not consistent with H_0
- Small values of X^2 suggests the data are consistent with H_0

χ^2 Test

In example 3 the contributions to the X^2 statistic are:

	Daily # cigarettes						
	None	< 5	5-14	15-24	25-49	50+	Total
Cancer	$\frac{(7-34)^2}{34}$	$\frac{(55-92)^2}{92}$	etc.				
Control	$\frac{(61-34)^2}{34}$						
Total							

	Daily # cigarettes						
	None	< 5	5-14	15-24	25-49	50+	Total
Cancer	21.44	14.88	3.10	1.07	21.61	6.76	
Control	21.44	14.88	3.10	1.07	21.61	6.76	
Total							

$$X^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} = 137.7$$

$p = P(X^2 > \chi^2(5) \mid H_0 \text{ true}) < 0.0001$

Conclusion?

χ^2 Test

	Factor Levels				
	1	2	…	C	Total
1	O_{11}	O_{12}	…	O_{1C}	N_1
Group 2	O_{21}				N_2
3	O_{31}				N_3
⋮	⋮				
R	O_{R1}			O_{RC}	N_R
Total	M_1	M_2		M_C	T

1. Compute the expected cell counts under homogeneity assumption:

   $$E_{ij} = N_i M_j / T$$

2. Compute the chi-square statistic:

   $$X^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

3. Compare X^2 to $\chi^2(df)$ where

   $$df = (R-1) \times (C-1)$$

4. Interpret acceptance/rejection or p-value.

2 x 2 Tables

Example 1: Pauling (1971)

Patients are randomized to either receive Vitamin C or placebo. Patients are followed-up to ascertain the development of a cold.

	Cold - Y	Cold - N	Total
Vitamin C	17	122	139
Placebo	31	109	140
Total	48	231	279

Q: Is treatment with Vitamin C associated with a reduced probability of getting a cold?

Q: If Vitamin C is associated with reducing colds, then what is the magnitude of the effect?

2 x 2 Tables

Example 2: Keller (AJPH, 1965)

Patients with (cases) and without (controls) oral cancer were surveyed regarding their smoking frequency (this table collapses over the smoking frequency categories).

	Case	Control	Total
Smoker	484	385	869
Non-Smoker	27	90	117
Total	511	475	986

Q: Is oral cancer associated with smoking?

Q: If smoking is associated with oral cancer, then what is the magnitude of the risk?

2 x 2 Tables

Example 3: Sex-linked traits

Suppose we collect a random sample of Drosophila and cross classify eye color and sex.

	male	female	Total
red	165	300	465
white	176	81	257
Total	341	381	722

Q: Is eye color associated with sex?

Q: If eye color is associated with sex, then what is the magnitude of the effect?

2 x 2 Tables

Example 4: Matched case control study

213 subjects with a history of acute myocardial infarction (AMI) were ***matched*** by age and sex with one of their siblings who did not have a history of AMI. The prevalence of a particular polymorphism was compared between the siblings

	AMI		
	carrier	noncarrier	Total
carrier	73	14	87
No AMI noncarrier	23	103	126
Total	96	117	213

Q: Is there an association between the polymorphism and AMI?

Q: If there is an association then what is the magnitude of the effect?

2 x 2 Tables

Each of these tables (except for example 4) can be represented as follows:

Disease Status

Exposure Status

	D	not D	Total
E	a	b	$(a + b) = n_1$
not E	c	d	$(c + d) = n_2$
Total	$(a + c) = m_1$	$(b + d) = m_2$	N

The question of association can be addressed with **Pearson's** X^2 (except for example 4) We compute the **expected** cell counts as follows:

Expected:

	D	not D	Total
E	n_1m_1/N	n_1m_2/N	$(a + b) = n_1$
not E	n_2m_1/N	n_2m_2/N	$(c + d) = n_2$
Total	$(a + c) = m_1$	$(b + d) = m_2$	N

2 x 2 Tables

Pearson's chi-square is given by:

$$X^2 = \sum_{i=1}^{4} (O_i - E_i)^2 / E_i$$

$$= \left(a - \frac{n_1m_1}{N}\right)^2 / \left(\frac{n_1m_1}{N}\right) + \left(b - \frac{n_1m_2}{N}\right)^2 / \left(\frac{n_1m_2}{N}\right) +$$

$$\left(c - \frac{n_2m_1}{N}\right)^2 / \left(\frac{n_2m_1}{N}\right) + \left(d - \frac{n_2m_2}{N}\right)^2 / \left(\frac{n_2m_2}{N}\right) +$$

$$= \frac{N(ad - bc)^2}{n_1n_2m_1m_2}$$

2 x 2 Tables

Example 1: Pauling (1971)

	Cold - Y	Cold - N	Total
Vitamin C	17 (12%)	122 (88%)	139
Placebo	31 (22%)	109 (78%)	140
Total	48	231	279

H_0 : probability of disease does not depend on treatment

H_A : probability of disease does depend on treatment

$$X^2 = \frac{N(ad-bc)^2}{n_1 n_2 m_1 m_2}$$

$$= \frac{279(17\times 109 - 31\times 122)^2}{139\times 140\times 48\times 231}$$

$$= 4.81$$

For the p-value we compute $P(\chi^2(1) > 4.81) = 0.028$. Therefore, we reject the homogeneity of disease probability in the two treatment groups.

2 x 2 Tables
Applications In Epidemiology

Example 1 fixed the number of E and not E, then evaluated the disease status after a fixed period of time (same for everyone). This is a **prospective study**. Given this design we can estimate the **relative risk**:

$$RR = \frac{P(D \mid E)}{P(D \mid \overline{E})}$$

The range of RR is [0, ∞). By taking the logarithm, we have (- ∞, +∞) as the range for ln(RR) and a better approximation to normality for the estimated $\ln(\hat{RR})$:

$$\ln(\hat{RR}) = \ln\left(\frac{\hat{P}(D \mid E)}{\hat{P}(D \mid \overline{E})}\right) = \ln\left(\frac{p_1}{p_2}\right)$$

$$= \ln\left(\frac{a/n_1}{c/n_2}\right)$$

$$\ln(\hat{RR}) \sim approx\ N\left(\ln(p_1/p_2), \frac{1-p_1}{p_1 n_1} + \frac{1-p_2}{p_2 n_2}\right)$$

Relative Risk

	Cold - Y	Cold - N	Total
Vitamin C	17	122	139
Placebo	31	109	140
Total	48	231	279

The estimated relative risk is:

$$\hat{RR} = \frac{\hat{P}(D \mid E)}{\hat{P}(D \mid \overline{E})}$$

$$= \frac{17/139}{31/140} = 0.55$$

We can obtain a 95% confidence interval for the relative risk by first obtaining a confidence interval for the log-RR:

$$\ln\left(\hat{RR}\right) \pm 1.96 \times \sqrt{\frac{1-p_1}{p_1 n_1} + \frac{1-p_2}{p_2 n_2}}$$

and exponentiating the endpoints of the CI.

Note that disease status and exposure status are transposed here compared to previous tables.

```
. csi 17 31 122 109

                |   Exposed   Unexposed  |      Total
----------------+------------------------+----------
          Cases |        17          31  |        48
       Noncases |       122         109  |       231
----------------+------------------------+----------
          Total |       139         140  |       279
                |                        |
           Risk |  .1223022    .2214286  |   .172043
                |                        |
                |      Point estimate    |  [95% Conf. Interval]
                |------------------------+----------------------
 Risk difference |        -.0991264       | -.1868592   -.0113937
      Risk ratio |         .5523323       |  .3209178    .9506203
 Prev. frac. ex. |         .4476677       |  .0493797    .6790822
 Prev. frac. pop |         .2230316       |
                +-------------------------------------------------
                             chi2(1) =     4.81  Pr>chi2 = 0.0283
```

2 x 2 Tables

Example 2: Keller (AJPH, 1965)

Patients with (cases) and without (controls) oral cancer were surveyed regarding their smoking frequency (this table collapses over the smoking frequency categories).

	Case	Control	Total
Smoker	484	385	869
Non-Smoker	27	90	117
Total	511	475	986

Q: Is oral cancer associated with smoking?

Q: If smoking is associated with oral cancer, then what is the magnitude of the risk?

2 x 2 Tables
Applications In Epidemiology

In **Example 2** we fixed the number of **cases** and **controls** then ascertained exposure status. Such a design is known as **case- control study**. Based on this we are able to directly estimate:

$$P(E \mid D) \quad \text{and} \quad P(E \mid \overline{D})$$

However, we generally are interested in the relative risk of disease given exposure, which is not estimable from these data alone - we've fixed the number of diseased and diseased free subjects, and it can be shown that in general:

$$P(D \mid E) \neq P(E \mid D)$$

$$\frac{P(D|E)}{P(D|\overline{E})} \neq \frac{P(E|D)}{P(E|\overline{D})}$$

Odds Ratio

Instead of the relative risk we can estimate the **exposure odds ratio** which (surprisingly) is equivalent to the **disease odds ratio**:

$$\frac{P(E \mid D)/(1-P(E \mid D))}{P(E \mid \overline{D})/(1-P(E \mid \overline{D}))} = \frac{P(D \mid E)/(1-P(D \mid E))}{P(D \mid \overline{E})/(1-P(D \mid \overline{E}))}$$

In other words, **the odds ratio can be estimated regardless of the sampling scheme.**

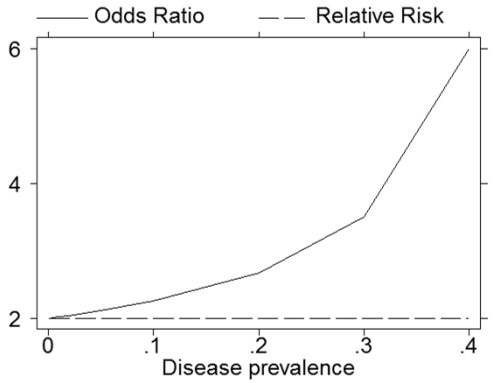
Furthermore, for rare diseases, P(D | E) ≈ 0 so that the disease odds ratio approximates the relative risk:

$$\frac{P(D \mid E)/(1-P(D \mid E))}{P(D \mid \overline{E})/(1-P(D \mid \overline{E}))} \approx \frac{P(D \mid E)}{P(D \mid \overline{E})}$$

Since with case-control data we are able to effectively estimate the exposure odds ratio we are then able to equivalently estimate the disease odds ratio which for rare diseases approximates the relative risk.

For rare diseases (e.g., prevalence <5%), the (sample) odds ratio estimates the (population) relative risk.

Odds Ratio



Odds Ratio

Like the relative risk, the odds ratio has [0, ∞) as its range. The **log odds ratio** has (- ∞, +∞) as its range and the normal approximation is better as an approximation to the dist of the estimated log odds ratio.

$$OR = \frac{p_1/1-p_1}{p_2/1-p_2}$$

$$\hat{OR} = \frac{\hat{p}_1/1-\hat{p}_1}{\hat{p}_2/1-\hat{p}_2}$$

$$\hat{OR} = \frac{ad}{bc}$$

Confidence intervals are based upon:

$$\ln\left(\hat{OR}\right) \sim N\left(\ln(OR), \frac{1}{n_1 p_1} + \frac{1}{n_1(1-p_1)} + \frac{1}{n_2 p_2} + \frac{1}{n_2(1-p_2)}\right)$$

Therefore, a 95% confidence interval for the log odds ratio is given by:

$$\ln\left(\frac{ad}{bc}\right) \pm 1.96 \times \sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}$$

Odds Ratio

```
. cci 484 27 385 90

                                                        Proportion
                |   Exposed   Unexposed  |      Total     Exposed
----------------+------------------------+------------------------
          Cases |       484          27  |        511      0.9472
       Controls |       385          90  |        475      0.8105
----------------+------------------------+------------------------
          Total |       869         117  |        986      0.8813
                |                        |
                |      Point estimate    |  [95% Conf. Interval]
                |------------------------+------------------------
     Odds ratio |         4.190476       |  2.633584    6.836229  (exact)
 Attr. frac. ex. |         .7613636       |  .6202893    .8537205  (exact)
 Attr. frac. pop |          .721135       |
                +-------------------------------------------------

                           chi2(1) =    43.95  Pr>chi2 = 0.0000
```

Interpreting Odds ratios

1. What is the outcome of interest? (i.e. disease)
2. What are the two groups being contrasted? (i.e. exposed and unexposed)

$$\text{OR} = \frac{\text{odds of OUTCOME in EXPOSED}}{\text{odds of OUTCOME in UNEXPOSED}}$$

- Similar to RR for rare diseases
- Meaningful for both cohort and case-control studies
- OR > 1 $\Rightarrow$ increased risk of OUTCOME with EXPOSURE
- OR < 1 $\Rightarrow$ decreased risk of OUTCOME with EXPOSURE

2 x 2 Tables

Example 3: Sex-linked traits

Suppose we collect a random sample of Drosophila and cross classify eye color and sex.

	male	female	Total
red	165	300	465
white	176	81	257
Total	341	381	722

Q: Is eye color associated with sex?

Q: If eye color is associated with sex, then what is the magnitude of the effect?

2 x 2 Tables
Applications in Epidemiology

Example 3 is an example of a **cross-sectional** study since only the total for the entire table is fixed in advance. The row totals or column totals are not fixed in advance.

	male	female	Total
red	165 (48%)	300 (79%)	465
white	176	81	257
Total	341	381	722

Cross-sectional studies

- Sample from the entire population, not by disease status or exposure status
- Use chi-square test to test for association
- Use RR or OR to summarize association
- Cases of disease are **prevalent** cases (compared to incident cases in a prospective or cohort study)

2 x 2 Tables
Applications in Epidemiology

Case = red eye color
Noncase = white eye color

```
                    male        female
-----------------+-----------------------+------------
           Cases |       165         300 |        465
       Noncases |       176          81 |        257
-----------------+-----------------------+------------
           Total |       341         381 |        722
            Risk |   .483871    .7874016 |   .6440443
                 |                       |
                 |      Point estimate   |    [95% Conf. Interval]
                 |-----------------------+------------------------
 Risk difference |        -.3035306      |   -.3706217   -.2364395
      Risk ratio |         .6145161      |     .544263    .6938375
 Prev. frac. ex. |         .3854839      |    .3061625     .455737
 Prev. frac. pop |         .1820637      |
      Odds ratio |          .253125      |    .1830613    .3500144
                 +-------------------------------------------------
                              chi2(1) =    72.32  Pr>chi2 = 0.0000
```

2 x 2 Tables

Example 4: Matched case control study

213 subjects with a history of acute myocardial infarction (AMI) were ***matched*** by age and sex with one of their siblings who did not have a history of AMI. The prevalence of a particular polymorphism was compared between the siblings

	AMI		
	carrier	noncarrier	Total
carrier	73	14	87
No AMI noncarrier	23	103	126
Total	96	117	213

Q: Is there an association between the polymorphism and AMI?

Q: If there is an association then what is the magnitude of the effect?

Paired Binary Data

Example 4 measures a binary response in sibs. This is an example of **paired binary data**. One way to display these data is the following:

	Carrier	Noncarrier	Total
AMI	96	117	213
No AMI	87	126	213
Total	183	243	426

Q: Can't we simply use X^2 Test of Homogeneity to assess whether this is evidence for an increase in knowledge?

A: NO!!! The X^2 tests assume that the rows are **independent** samples. In this design the 213 with AMI are genetically related to the 213 w/o AMI.

Paired Binary Data

For paired binary data we display the results as follows:

	AMI	
	1	0
No AMI 1	n_{11}	n_{10}
0	n_{01}	n_{00}

This analysis explicitly recognizes the heterogeneity of subjects. Thus, those that score (0,0) and (1,1) provide no information about the association between AMI and the polymorphism. These are known as the **concordant pairs**. The information regarding the association is in the **discordant pairs**, (0,1) and (1,0).

p_1 = P(carrier | AMI)

p_0 = P(carrier | No AMI)

$H_0 : p_1 = p_0$

$H_A : p_1 \neq p_0$

$$\hat{p}_1 - \hat{p}_0 = \frac{n_{11} + n_{01}}{N} - \frac{n_{11} + n_{10}}{N} = \frac{n_{01} - n_{10}}{N}$$

Paired Binary Data
McNemar's Test

Under the null hypothesis, $H_0 : p_1 = p_0$, we expect equal numbers of 01's and 10's. ($E[n_{01}] = E[n_{10}]$). Specifically, under the null:

$$M = n_{01} + n_{10}$$

$$n_{10} \mid M \sim Bin\left(M, \frac{1}{2}\right)$$

$$Z = \frac{n_{10} - M\frac{1}{2}}{\sqrt{M\frac{1}{2}\left(1 - \frac{1}{2}\right)}}$$

Under H_0, $Z^2 \sim \chi^2(1)$, and forms the basis for **McNemar's Test for Paired Binary Responses**.

The odds ratio comparing the odds of carrier in those with AMI to odds of carrier in those w/o AMI is estimated by:

$$\hat{OR} = \frac{n_{01}}{n_{10}}$$

Confidence intervals can be obtained as described in Breslow and Day (1981), section 5.2, or in Armitage and Berry (1987), chapter 16.

Example 4:

	AMI		
	carrier	noncarrier	Total
carrier	73	14	87
No AMI noncarrier	23	103	126
Total	96	117	213

We can test H_0: $p_1 = p_2$ using **McNemar's Test:**

$$Z = \frac{n_{01} - M_{\frac{1}{2}}^{1}}{\sqrt{M_{\frac{1}{2}}^{1}\left(\frac{1}{2}\right)}}$$

$$= \frac{23 - (23+14)/2}{\sqrt{(23+14)/4}}$$

$$= 1.48$$

Comparing 1.48^2 to a χ^2 (1) we find that $p > 0.05$. Therefore, we do not reject the null hypothesis and find little evidence of association between gene and disease.

We estimate the odds ratio as $\hat{OR} = 23/14 = 1.64.$

Matched case-control data

```
. mcci 73 23 14 103

                 | Controls                |
Cases            |   Exposed   Unexposed   |      Total
-----------------+-------------------------+-----------
         Exposed |          73          23 |         96
       Unexposed |          14         103 |        117
-----------------+-------------------------+-----------
           Total |          87         126 |        213

McNemar's chi2(1) =      2.19    Prob > chi2 = 0.1390
Exact McNemar significance probability       = 0.1877

Proportion with factor
        Cases          .4507042
        Controls       .4084507     [95% Conf. Interval]
                       ---------    --------------------
        difference     .0422535     -.0181247   .1026318
        ratio          1.103448      .9684942   1.257207
        rel. diff.     .0714286     -.0197486   .1626057

        odds ratio     1.642857      .8101776   3.452833   (exact)
```

Two way tables - Review

- How were data collected?
  - Cohort design
  - Case-control design
  - Cross-sectional design
  - Matched pairs
- Is there an association?
  - R x C Tables
    - Chi-square tests of Homogeneity & Independence
  - 2 x 2 Tables
    - Chi-square test
    - Paired data and McNemar's
- What is the magnitude of the association?
  - Relative risk
  - Odds ratio (≈ relative risk for rare diseases)
  - Risk difference (attributable risk)

SUMMARY
Measures of Association for 2 x 2 Tables

RD = $p_1 - p_2$ = risk difference (null: RD = 0)

- also known as **attributable risk** or **excess risk**
- measures **absolute effect** – the proportion of cases among the exposed that can be attributed to exposure

RR = p_1 / p_2 = relative risk (null: RR = 1)

- measures **relative effect** of exposure
- bounded above by $1/p_2$

OR = $[p_1(1-p_2)]/[p_2(1-p_1)]$ = odds ratio (null: OR = 1)

- range is 0 to ∞
- approximates RR for rare events
- invariant of switching rows and cols
- good behavior of p-values and CI even for small to moderate sample size

SUMMARY
Models for 2 x 2 Tables

1. **Cohort** ("Prospective", "Followup")
   - Sample n_1 "exposed" and n_2 "unexposed"
   - Follow everyone for equal period of time
   - Observe incident disease – r_1 cases among exposed, r_2 cases among unexposed
   - Model: Two independent binomials

     $r_1 \sim \text{binom}(p_1, n_1)$

     $r_2 \sim \text{binom}(p_2, n_2)$

     $p_1 = P(D|E)$

     $p_2 = P(D|\overline{E})$
   - Useful measures of association – RR,OR,RD
   - Examples:

     r_i = number of cases of HIV during 1 year followup of n_i individuals in arm i of HIV prevention trial

     r_i = number of low birthweight babies among n_i live births

SUMMARY
Models for 2 x 2 Tables

2. **Case-Control**
   - Sample n_1 "cases" and n_2 "controls"
   - Observe exposure history – r_1 exposed among cases, r_2 exposed among controls
   - Model: Two independent binomials

     $r_1 \sim \text{binom}(q_1, n_1)$

     $r_2 \sim \text{binom}(q_2, n_2)$

     $q_1 = P(E|D)$

     $q_2 = P(E|\overline{D})$
   - Useful measures of association – OR
   - Examples:

     r_i = consistent condom use (yes/no) among those with/without HPV infection

     r_i = number exposed to alcohol during pregnancy among n_i low birthweight/normal birthweight babies

SUMMARY
Models for 2 x 2 Tables

3. Cross-sectional

- Sample n individuals from population
- Observe both "exposure" and (prevalent) "disease" status.
- No longitudinal followup
- Useful measures of association – RR,OR,RD
- Example:

  n_{ij} = number of gay men with gonorrhea in random sample of STD clinic attendees

Stratified Tables

- Often, a third measure influences the relationship between the two primary measures (i.e. disease and exposure).
- How do we "remove or control for the effect" of the third measure?
- Issues of causality

Example: Effect of seat belt use on accident fatality

	Seat Belt	
Driver	Worn	Not worn
dead	10	20
alive	40	30
Total	50	50
Fatality Rate	10/50 (20%)	20/50 (40%)

Stratified Tables

But, suppose...

	Impact Speed			
	≤ 40 mph		> 40 mph	
Driver	seat belt		seat belt	
	worn	not	worn	not
dead	3	2	7	18
alive	27	18	13	12
Total	30	20	20	30
Fatality Rate	10%	10%	35%	60%

How does this affect your inference?

- This is an example of "effect modification" or "interaction".

Stratified tables - Confounding (Simpson's Paradox)

Differences in surgical success between hospitals?

		Death rate	
Hospital	A	63/2100	(3%)
	B	16/800	(2%)

BUT ...

		Death rate	
High risk			
Hospital	A	57/1500	(3.8%)
	B	8/200	(4%)
Low risk			
Hospital	A	6/600	(1%)
	B	8/600	(1.3%)

Explanation: Higher risk individuals are more likely to die AND are more likely to go to hospital A (perhaps it specializes in this type of surgery)

Confounding

"A confounding variable is a variable that is associated with both the disease and the exposure variable." *Rosner (1995)*

"Confounding is the distortion of a disease/exposure association brought about by the association of other factors with both disease and exposure, the latter associations with disease being causal." *Breslow & Day (1980)*

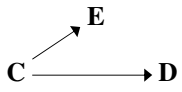
"If any factor either increasing or decreasing the risk of a disease besides the characteristic or exposure under study is unequally distributed in the groups that are being compared with regard to the disease, this itself will give rise to differences in disease frequency in the compared groups. Such distortion, termed confounding, leads to an invalid comparison." *Lilienfeld & Stolley (1994)*

Confounding

A confounder is associated with both the disease and exposure and is not in the causal path between disease and exposure

- The implicit assumption is that we want to know if E "causes" D
- A simple, common example from genetics is the linked gene: we discover a gene which appears to be associated with disease … does it cause the disease or is it merely linked to the true causal gene?

Pictorially …



An apparent association between E and D is completely explained by C. C is a confounder.

Adjusting the OR via Stratification

Basic idea

- Compute separate OR for each stratum
- Assess homogeneity of OR's across strata
- Pool OR's: used weighted average
- Global test of pooled OR = 1
- Different methods of pooling, testing have been proposed. We will focus on Mantel-Haenszel methods
- Same idea for RR and RD

Stratified Contingency Tables - Example

EXAMPLE:

Suppose we are interested in the relationship between lung-cancer incidence and heavy drinking (defined as ≥ 2 drinks per day). We conduct a prospective study where drinking status is determined at baseline and the cohort is followed for 10 years to determine cancer endpoints. We also measure smoking status at baseline.

Stratified Contingency Tables - Example

1) Pooled data, not controlling for smoking

	Heavy Drinker		
	Yes	No	
Case	33	27	60
Control	1667	2273	3940
	1700	2300	4000

```
. cci 33 1667 27 2273
                                                        Proportion
                |   Exposed   Unexposed  |      Total     Exposed
----------------+------------------------+------------------------
          Cases |        33        1667  |       1700      0.0194
       Controls |        27        2273  |       2300      0.0117
----------------+------------------------+------------------------
          Total |        60        3940  |       4000      0.0150
                |                        |
                |      Point estimate    |  [95% Conf. Interval]
                |------------------------+------------------------
     Odds ratio |         1.666533       |   .9677794    2.892949  (exact)
 Attr. frac. ex. |          .399952       |  -.0332933    .6543319  (exact)
 Attr. frac. pop |         .0077638       |
                +-------------------------------------------------
                           chi2(1) =     3.89  Pr>chi2 = 0.0484
```

Stratified Contingency Tables - Example

2) Stratified by smoking at baseline

Smokers

	Heavy Drinking		
	Yes	No	
Case	24	6	30
Control	776	194	970
	800	200	1000

```
. cci 24 6 776 194

                                                        Proportion
                |   Exposed   Unexposed  |      Total     Exposed
----------------+------------------------+------------------------
          Cases |        24           6  |       30      0.8000
       Controls |       776         194  |      970      0.8000
----------------+------------------------+------------------------
          Total |       800         200  |     1000      0.8000
                |                        |
                |      Point estimate    |  [95% Conf. Interval]
                |------------------------+------------------------
     Odds ratio |                 1      |  .3911965    3.033018  (exact)
 Attr. frac. ex. |                 0      |  -1.55626    .6702954  (exact)
 Attr. frac. pop |                 0      |
                +-------------------------------------------------
                           chi2(1) =     0.00  Pr>chi2 = 1.0000
```


Stratified Contingency Tables - Example

Nonsmokers

	Heavy Drinking		
	Yes	No	
Case	9	21	30
Control	891	2079	2970
	900	2100	3000

```
. cci 9 21 891 2079

                                                        Proportion
                |   Exposed   Unexposed  |      Total     Exposed
----------------+------------------------+------------------------
          Cases |         9          21  |       30      0.3000
       Controls |       891        2079  |     2970      0.3000
----------------+------------------------+------------------------
          Total |       900        2100  |     3000      0.3000
                |                        |
                |      Point estimate    |  [95% Conf. Interval]
                |------------------------+------------------------
     Odds ratio |                 1      |  .4015748    2.288393  (exact)
 Attr. frac. ex. |                 0      | -1.490196    .5630121  (exact)
 Attr. frac. pop |                 0      |
                +-------------------------------------------------
                           chi2(1) =     0.00  Pr>chi2 = 1.0000
```

Stratified Contingency Tables

Q: How can we combine the information from both tables to obtain an overall test of significance that takes account of the stratification?

A: Mantel-Haenszel Methods – assesses association between disease and exposure after controlling for one or more confounding variables.

Notation:

	E	$\overline{E}$	
D	a_i	b_i	$(a_i + b_i)$
$\overline{D}$	c_i	d_i	$(c_i + d_i)$
	$(a_i + c_i)$	$(b_i + d_i)$	N_i

where i = 1,2,…,K is the number of strata.

Mantel-Haenszel Methods

(1) **Test of effect modification** (heterogeneity, interaction)

Ho: $OR_1 = OR_2 = \ldots = OR_K$

Ha: not all stratum-specific OR's are equal

(2) **Estimate the common odds ratio**

The Mantel-Haenszel estimate of the odds ratio assumes there is a **common** odds ratio:

$$OR_{pool} = OR_1 = OR_2 = \ldots = OR_K$$

To estimate the common odds ratio we take a weighted average of the stratum-specific odds ratios:

MH estimate: $\hat{OR}_{pool} = \sum_{i=1}^{K} w_i \cdot \hat{OR}_i$

(3) **Test of common odds ratio**

H_o: common odds ratio is 1.0

H_a: common odds ratio ≠ 1.0

Mantel-Haenszel Methods - Example

Lung Cancer data

```
. use "P:\Biostat513_06\drink.dta", clear
. list

     +----------------------------------+
     | cancer   drink   number   smoke |
     |----------------------------------|
  1. |      1       1       24       1 |
  2. |      1       0        6       1 |
  3. |      0       1      776       1 |
  4. |      0       0      194       1 |
  5. |      1       1        9       0 |
  6. |      1       0       21       0 |
  7. |      0       1      891       0 |
  8. |      0       0     2079       0 |
     +----------------------------------+

. cc cancer drink [freq=number], by(smoke) bd

          Smoker |       OR      [95% Conf. Interval]    M-H Weight
-----------------+-------------------------------------------------
               0 |          1     .4015748   2.288393         6.237 (exact)
               1 |          1     .3911965   3.033018         4.656 (exact)
-----------------+-------------------------------------------------
           Crude |   1.666533     .9677794   2.892949               (exact)
    M-H combined |          1     .5521991   1.810941
---------------------------------------------------------------------
Test of homogeneity (M-H)      chi2(1) =     0.00  Pr>chi2 = 1.0000
Test of homogeneity (B-D)      chi2(1) =     0.00  Pr>chi2 = 1.0000

                  Test that combined OR = 1:
                              Mantel-Haenszel chi2(1) =      0.00
                                       Pr>chi2 =    1.0000
```


Stratified Contingency Tables - Example

EXAMPLE: (Rosner sec 13.5)

A 1985 study identified a group of 518 cancer cases and a group of age- and sex-matched controls by mail questionnaire. The main purpose of the study was to look at the effect of passive smoking on cancer risk. In the study passive smoking was defined as exposure to the cigarette smoke of a spouse who smoked at least one cigarette/day for at least 6 months. One potential confounding variable was smoking by the test subjects themselves since personal smoking is related to both cancer risk and having a spouse that smokes. Therefore, it was important to control for personal smoking before looking at the relationship between passive smoking and cancer risk.

Stratified Contingency Tables - Example

1) Pooled data, not controlling for personal smoking

	Passive smoking		
	Yes	No	
Case	281	228	509
Control	210	279	489
	491	507	998

```
. cci 281 228 210 279

                                                      Proportion
                |   Exposed   Unexposed  |      Total     Exposed
----------------+------------------------+------------------------
          Cases |       281         228  |        509      0.5521
       Controls |       210         279  |        489      0.4294
----------------+------------------------+------------------------
          Total |       491         507  |        998      0.4920
                |                        |
                |      Point estimate    |  [95% Conf. Interval]
                |------------------------+------------------------
     Odds ratio |          1.637406      |  1.265013    2.119599  (exact)
 Attr. frac. ex. |          .3892779      |  .2094943    .5282126  (exact)
 Attr. frac. pop |          .2149059      |
                +-------------------------------------------------
                           chi2(1) =    15.00  Pr>chi2 = 0.0001
```


Stratified Contingency Tables - Example

2) Stratified by personal smoking

Nonsmokers

	Passive smoking		
	Yes	No	
Case	120	111	231
Control	80	155	235
	200	266	466

```
. cci 120 111 80 155

                                                      Proportion
                |   Exposed   Unexposed  |      Total     Exposed
----------------+------------------------+------------------------
          Cases |       120         111  |        231      0.5195
       Controls |        80         155  |        235      0.3404
----------------+------------------------+------------------------
          Total |       200         266  |        466      0.4292
                |                        |
                |      Point estimate    |  [95% Conf. Interval]
                |------------------------+------------------------
     Odds ratio |          2.094595      |   1.41754    3.097165  (exact)
 Attr. frac. ex. |          .5225806      |  .2945527    .6771241  (exact)
 Attr. frac. pop |          .2714705      |
                +-------------------------------------------------
                           chi2(1) =    15.24  Pr>chi2 = 0.0001
```

Stratified Contingency Tables - Example

Smokers

	Passive smoking		
	Yes	No	
Case	161	117	278
Control	130	124	254
	291	241	532

```
. cci 161 117 130 124
                                                    Proportion
                |   Exposed   Unexposed  |      Total     Exposed
----------------+------------------------+------------------------
          Cases |       161         117  |       278      0.5791
       Controls |       130         124  |       254      0.5118
----------------+------------------------+------------------------
          Total |       291         241  |       532      0.5470
                |                        |
                |     Point estimate     |  [95% Conf. Interval]
                |------------------------+------------------------
     Odds ratio |         1.312558       |  .9184614    1.875813  (exact)
 Attr. frac. ex. |         .2381286       | -.0887774    .4668978  (exact)
 Attr. frac. pop |          .137909       |
                +-------------------------------------------------
                           chi2(1) =     2.43  Pr>chi2 = 0.1192
```


Mantel-Haenszel Methods - Example

Passive Smoking data

```
. use "M:\.MyDocs\b513\passive.dta"
. list

     +---------------------------------+
     | case   passive   number   smoke |
     |---------------------------------|
  1. |    1         1      120       0 |
  2. |    1         0      111       0 |
  3. |    0         1       80       0 |
  4. |    0         0      155       0 |
  5. |    1         1      161       1 |
  6. |    1         0      117       1 |
  7. |    0         1      130       1 |
  8. |    0         0      124       1 |
     +---------------------------------+

. cc case passive [freq=number], by(smoke) bd

Personal Smoking |       OR       [95% Conf. Interval]    M-H Weight
-----------------+-------------------------------------------------
               0 |   2.094595      1.41754   3.097165      19.05579 (exact)
               1 |   1.312558     .9184614   1.875813      28.59023 (exact)
-----------------+-------------------------------------------------
           Crude |   1.637406     1.265013   2.119599               (exact)
    M-H combined |   1.625329     1.263955   2.090024
-------------------------------------------------------------------
Test of homogeneity (M-H)       chi2(1) =     3.27  Pr>chi2 = 0.0706
Test of homogeneity (B-D)       chi2(1) =     3.27  Pr>chi2 = 0.0704

                    Test that combined OR = 1:
                                Mantel-Haenszel chi2(1) =     14.42
                                                 Pr>chi2 =    0.0001
```

Stratified Data - Summary

1. Compute stratum-specific measures
2. Evaluate stratum-specific estimates by a test of homogeneity. Consider test results in light of sample size.
3. If the homogeneity test result is non-significant then consider a common estimate, pooling across all strata
   (a) calculate an overall (common) summary (OR)
   (b) test for significant association
   (c) calculate confidence interval
4. If the homogeneity test result is significant then we are concerned that the ORs vary across strata. We may
   (a) If the direction of association (+) is same and the difference is small in magnitude, then
      - proceed as in 3 above (calculating average summary)
      - report on the test of homogeneity.
   (b) If the direction of the association is different, then
      - report results from test of homogeneity
      - report stratum-specific measures and confidence intervals.
      - does the average make sense at all?

Review

- R x C contingency table
  - Test for homogeneity (Pearson chi-squared)
- Single 2 x 2 table
  - Different sampling schemes
    1. Cohort (row totals fixed)
    2. Case-control (column totals fixed)
    3. Cross-sectional (grand total fixed)
  - Different measures of association
    RD (Designs 1 & 3)
    RR (Designs 1 & 3)
    OR (Designs 1, 2 & 3)
  - Test of association
    Pearson chi-squared
    McNemar's
    Fisher exact

Review

- Series of 2 x 2 tables
  - o Mantel-Haenszel (combined) OR estimate
  - o Mantel-Haenszel test for association

    H_o: OR = 1

    H_a: OR constant, $\neq 1$
  - o Breslow-Day "Score" Test for Homogeneity (Interaction, Effect Modification)

The Bootstrap and Jackknife

Bootstrap & Jackknife Motivation

In scientific research

- Interest often focuses upon the estimation of some unknown parameter, θ. The parameter θ can represent for example, mean weight of a certain strain of mice, heritability index, a genetic component of variation, a mutation rate, etc.
- Two key questions need to be addressed:
  1. How do we estimate θ?
  2. Given an estimator for θ, how do we estimate its precision/accuracy?
- We assume Question 1 can be reasonably well specified by the researcher
- Question 2, for our purposes, will be addressed via the estimation of the estimator's standard error

Bootstrap Motivation

Challenges

• Answering Question 2, even for relatively simple estimators (e.g., ratios and other non-linear functions of estimators) can be quite challenging

  • Solutions to most estimators are mathematically intractable or too complicated to develop (with or without advanced training in statistical inference)

• However

  • Great strides in computing, particularly in the last 25 years, have made computational intensive calculations feasible.

• We will investigate how the bootstrap allows us to obtain robust estimates of precision for our estimator, θ, with a simple example…

Bootstrap Estimation

Estimating the precision of the sample mean

• A dataset of n observation provides more than an estimate of the population mean (denoted here as $\overline{X}$), where

  • $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$

• It gives an estimate of the precision of $\overline{X}$, namely

  • $\hat{se}\left[\overline{X}\right] = \sqrt{\hat{\sigma}^2 / n}$,

  where $\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} (X_i - \overline{X})^2}{n-1}$,

  is an estimate of the population variance.

• The problem with this standard error estimate is that it does not extend to estimators other than $\overline{X}$ in an obvious way.

Bootstrap Estimation

Estimating the precision of the sample mean

• From the formulas on the previous page, we can obtain an estimate of precision for $\overline{X}$ by estimating the population variance and "plugging" it into the formula for the standard error estimate.

• **Question**: What IF you did not know the formula for the standard error of the sample mean, BUT you had access to modern PC. How might you obtain an estimate of precision?

• **Answer**: The bootstrap!

Bootstrap Algorithm

Bootstrapping

• Assuming the sample accurately reflects the population from which it is drawn $\overline{X}$

• Generate a large number of "bootstrap" samples by resampling (with replacement) from the dataset

• Resample with the same structure (dependence, sample sizes) as used in the original sample

• Compute your estimator, $\hat{\theta}$, (here, $\hat{\theta} = \overline{X}$), for each of the bootstrap samples

• Compute the "standard deviation" from the statistics calculated above.

Bootstrap Algorithm

Bootstrap sample | Bootstrap estimates

1: $(X_1^{(1)}, X_2^{(1)}, \ldots, X_n^{(1)}) \rightarrow \hat{\theta}(X_1^{(1)}, X_2^{(1)}, \ldots, X_n^{(1)}) = \overline{X}^{(1)}$

2: $(X_1^{(2)}, X_2^{(2)}, \ldots, X_n^{(2)}) \rightarrow \hat{\theta}(X_1^{(2)}, X_2^{(2)}, \ldots, X_n^{(2)}) = \overline{X}^{(2)}$

$\vdots$ $\vdots$

B: $(X_1^{(B)}, X_2^{(B)}, \ldots, X_n^{(B)}) \rightarrow \hat{\theta}(X_1^{(B)}, X_2^{(B)}, \ldots, X_n^{(B)}) = \overline{X}^{(B)}$

Compute $\hat{\sigma}_b^2$, where $\hat{\sigma}_b^2 = \dfrac{\sum_{j=1}^{B} (\overline{X}^{(j)} - \overline{\overline{X}}^{(\cdot)})^2}{B-1}$, and $\overline{\overline{X}}^{(\cdot)} = \dfrac{1}{B}\sum_{j=1}^{B} \overline{X}^{(j)}$.

The bootstrap standard error is $\hat{se}_{boot}\left[\overline{X}\right] = \sqrt{\hat{\sigma}_b^2}$.

For other estimators, simply replace $\overline{X}$ with the $\hat{\theta}$ of your choice.

Bootstrap Estimation Examples

Estimating the precision of the sample mean

• Example: Generated a sample of size n=49 observations with the following summary statistics:

- $\overline{X} = \frac{1}{49}\sum_{i=1}^{n} X_i = 49.71$
- $\hat{se}\left[\overline{X}\right] = \sqrt{\hat{\sigma}^2 / n} = \sqrt{49.104/49} = 1.001$

• We generated B=100,000 bootstrap samples of size n=49 to obtain 100,000 bootstrap estimates of the sample mean, i.e., $\overline{X}^{(1)}, \overline{X}^{(2)}, \ldots, \overline{X}^{(100,000)}$.

• The bootstrap standard error was

- $\hat{se}_{boot}\left[\overline{X}\right] = \sqrt{\hat{\sigma}_b^2} = \sqrt{0.982} = 0.991$

• A reasonably close estimate to the "true" standard error estimate of 1.001

Bootstrap Estimation Examples

Confidence Intervals on the Sample Median

• Approximate confidence intervals for the median can be obtained using asymptotic theory

- The sample median is asymptotically normally distributed
- The formula for the standard error is difficult to use

$$X_m \sim N\left(mdn(X), \frac{1}{4\left[f(mdn(X))\right]^2}\right)$$

where f is the density function of the true median.

•Approximate confidence intervals for the median can be obtained using asymptotic theory

• Bootstrapping would be easier/easiest.

Bootstrap Estimation Examples

Bootstrapped estimates of the standard error for sample median

	Data	Median
Original sample:	{1, 5, 8, 3, 7}	5
Bootstrap 1:	{1, 7, 1, 3, 7}	3
Bootstrap 2:	{7, 3, 8, 8, 3}	7
Bootstrap 3:	{7, 3, 8, 8, 3}	7
Bootstrap 4:	{3, 5, 5, 1, 5}	5
Bootstrap 5:	{1, 1, 5, 1, 8}	1
etc.		
Bootstrap B (=1000)		

Bootstrap Estimation Examples

Bootstrapped estimates of the standard error for sample median (cont.)

- Descriptive statistics for the sample medians from 1000 bootstrap samples

B	1000
Mean	4.964
Standard Deviation	**1.914**
Median	5
Minimum, Maximum	1, 8
25th, 75th percentile	3, 7

- We estimate the standard error for the sample median as 1.914

- A 95% asymptotic (with n=5?) confidence interval (using the 0.975 quantile of the standard normal distribution) is

5 +/- 1.96(1.914) = (1.25, 8.75)

Bootstrap Estimation Examples

Confidence Intervals on the relative risk

- Approximate confidence intervals for the estimated relative risk, r = P[D|Exposed]/P[D|Not exposed] can also be obtained using asymptotic theory

  - The log[r] is asymptotically normally distributed with mean equal to the log of the true relative risk and variance

$$\text{var}\left[\log(r)\right] = \frac{1-P[D\,|\,E]}{n_1 P[D\,|\,E]} + \frac{1-P[D\,|\,\overline{E}]}{n_2 P[D\,|\,\overline{E}]}$$

  - 95% confidence intervals for the relative risk are therefore obtained by using the 0.975 quantile of the standard normal distribution (1.96) in the formula

$$\left(r\times\exp\left[-\sqrt{1.96\,\text{var}\left[\log(r)\right]}\right], r\times\exp\left[+\sqrt{1.96\,\text{var}\left[\log(r)\right]}\right]\right)$$

- We'll compare this approximation to the bootstrap in our example below

Bootstrap Estimation Examples

Bootstrapped estimates of the standard error for sample relative risk

Cross-classification of Framingham Men by high systolic blood pressure and heart disease

	Heart Disease	
High Systol BP	No	Yes
No	915	48
Yes	322	44

The sample estimate of the relative risk is

$r = (44/366)/(48/963) = 2.412$

The asymptotic 95% confidence interval is

(2.412*0.756, 2.412*1.322) = (1.82, 3.19).

Bootstrap Estimation Examples

Bootstrapped estimates of the standard error for the relative risk (cont.)

- Descriptive statistics for the sample relative risks

B	100000
Bootstrap mean, *r*	2.464
Bootstrap Median	2.412
Standard Deviation	**0.507**

- The bootstrap standard error for the estimated relative risk is 0.507
- A 95% bootstrap confidence interval is

2.412 +/- 1.96(0.507) = (1.42, 3.41)

Bootstrap Summary

Advantages

• All purpose computer intensive method useful for statistical inference.
• Bootstrap estimates of precision do not require knowledge of the theoretical form of an estimator's standard error, no matter how complicated it is.

Disadvantages

• Typically not useful for correlated (dependent) data.
• Missing data, censoring, data with outliers are also problematic.

Jackknife

Jackknife Estimation

• The jackknife (or leave one out) method, invented by Quenouille (1949), is an alternative resampling method to the bootstrap.

• The method is based upon sequentially deleting one observation from the dataset, recomputing the estimator, here, $\hat{\theta}_{(i)}$, n times. That is, there are exactly n jackknife estimates obtained in a sample of size n.

• Like the bootstrap, the jackknife method provides a relatively easy way to estimate the precision of an estimator, θ.

• The jackknife is generally less computationally intensive than the bootstrap

Jackknife Algorithm

Jackknifing

• For a dataset with n observations, compute n estimates by sequentally omitting each observation from the dataset and estimating $\hat{\theta}$ on the remaining $n-1$ observations.

• Using the n jackknife estimates, $\hat{\theta}_{(1)}, \hat{\theta}_{(2)}, \ldots, \hat{\theta}_{(n)}$,

we estimate the standard error of the estimator as

$$\hat{se}_{jack} = \sqrt{\frac{n-1}{n}\sum_{i=1}^{n}(\hat{\theta}_{(i)} - \bar{\hat{\theta}}_{(\cdot)})^2}$$

• Unlike the bootstrap, the jackknife standard error estimate will not change for a given sample

Jackknife Summary

Advantages

• Useful method for estimating and compensating for bias in an estimator.
• Like the bootstrap, the methodology does not require knowledge of the theoretical form of an estimator's standard error.
• Is generally less computationally intensive compared to the bootstrap method.

Disadvantages

• The jackknife method is more conservative than the bootstrap method, that is, its estimated standard error tends to be slightly larger.
• Performs poorly when the the estimator is not sufficiently smooth, i.e., a non-smooth statistics for which the jackknife performs poorly is the median.

Permutation and Exact Tests

&

False Detection Rate

Exact and Permutation Tests

- Computer-intensive methods for hypothesis testing
- Permutation Test (randomization test): Used when distribution of the test statistic (under the null hypothesis) is unknown
- Exact Test: Used when sample sizes are small, so standard asymptotic (large sample) procedures do not work well
- All permutation tests are exact tests but not vice-versa. Exact test maintains the Type I error level without any large sample approximations/assumptions

Example - HPV vaccine trial

- 200 uninfected women are randomly assigned 1:1 to HPV vaccine or placebo (i.e., 100 to each group)
- After 1 year subjects are tested for HPV infection (yes/no)
- Does the probability of infection differ between the two groups?

What is a useful model for these data?

Example - HPV vaccine trial

Vaccine group: Binomial(100, p_V)

Placebo group: Binomial(100, p_P)

Scientific Question:

Is the risk of infection the same or different in the two groups?

Restate in terms of the model:

H_0: $p_v = p_p$ ("null hypothesis")

vs. H_a: $p_v < p_p$

Example - HPV vaccine trial

Results:

	Vaccine	Placebo	Total
HPV+	20	40	60
HPV-	80	60	140
	100	100	200

The overall infection rate is 30%, but we observe 20% and 40% for vaccine and placebo, respectively. What if we repeated the experiment … would we see similar results? We know that sample results are variable. Could the difference go the other way? Could a difference this large be due to chance alone?

Example - HPV vaccine trial

We first need a way of summarizing the difference in the infection probabilities between vaccine and placebo groups. A useful summary has these features:

- Summarize the differences between the groups in a single number.

Example $\Rightarrow$ $p_v - p_p$

- One particular value (say, 0) of the summary corresponds to the null hypothesis being exactly true.

Example $\Rightarrow$ $p_v - p_p = 0$

- We expect values near 0 if the null hypothesis is true; we expect values far from 0 if the null hypothesis is false.
- But how near is near? How far is far?

Example - HPV vaccine trial

We need to figure out what sort of distribution of values we would see for our summary statistic if the experiment were repeated many times and the null hypothesis were true.

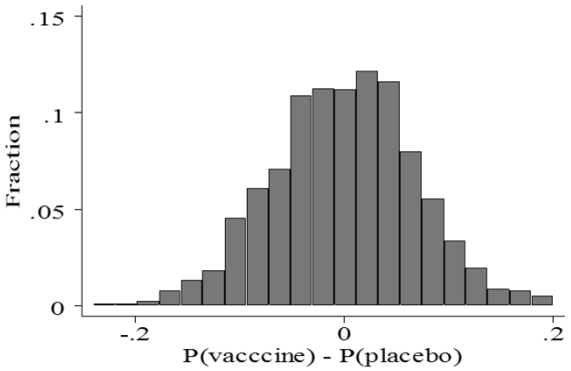
Imagine the following experiment:

- make up a deck of 200 cards
- mark the word “HPV+" on 60 of them
- shuffle and deal two groups of 100
- form a 2 x 2 table from the results
- calculate your summary statistic
- repeat many times
- plot the results

This experiment should give us an idea of what we expect to see **if the null hypothesis is true**.

Example - HPV vaccine trial

Here is the distribution of differences $p_v - p_p$ that we might expect to see if the null hypothesis is true:



Summarize the results by reporting what proportion of the simulated results are as “extreme” or more so than the observed result (p value).

⇒ **only 3/2000 simulated differences were more extreme than the observed difference of -0.2**

⇒ **p = .0015**

Example - HPV vaccine trial

Summary:

We have constructed a valid test of the hypothesis, H_0: $p_V = p_P$, using a **randomization test**. There are four steps involved:

1. Pick a model for the data and restate the scientific question in terms of the model (null hypothesis)
2. Choose (any) reasonable summary statistic that quantifies deviations from the null hypothesis
3. Resample data assuming the null hypothesis is true and compute the summary statistic for each resampled data set.
4. Compare the observed value of the summary statistic to the null distribution generated in Step 3.

Permutation Test for Correlation

Assume data are pairs (X_{1i},X_{2i}), i = 1,2,…,n

1. Ho: $\rho = 0$
2. Compute $r_{obs} = \text{corr}\ (X_1,X_2)$
3. Mix up the X_{1i} and X_{2i} ; i.e., for each X_{1i} randomly choose X_{2i}' from all the X_2's. Compute $r_{perm} = \text{corr}(X_1,X_2')$
4. Repeat Step 3 many times and compare r_{obs} to the distribution of r_{perm}

Note: There are n! possible pairings. If n is small, you can enumerate all possible pairings.

Permutation Tests - Summary

- Useful when we can do resampling under the null hypothesis
- Permutation samples are drawn without replacement
- If the sample size is small, you can enumerate all possible permutations, otherwise generate many permutations.
- Fewer assumptions than e.g. t-test (i.e., no assumption about skewness or normality of underlying distribution)
- Many standard nonparametric methods (e.g., Wilcoxon Rank Sum Test) are permutation tests based on ranks.
- Good Reference:
  Manly (2007). *Randomization, Bootstrap and Monte Carlo Methods in Biology*. Chapman & Hall/CRC.

Fisher's Exact Test

Motivation: When a 2×2 table contains cells that have fewer than 5 expected observations, the χ^2 approximation to the distribution of X^2 is known to be poor. This can lead to incorrect inference since the p-values based on this approximation are not valid.

Solution: Use Fisher's Exact Test

	D+	D-	Total
E+	*a*	*b*	n_1
E-	*c*	*d*	n_2
Total	m_1	m_2	N

Fisher's Exact Test

Example: A retrospective study is done among men aged 50-54 who died over a 1-month period. The investigators tried to include equal numbers of men who died from cardiovascular disease (CVD) and those that did not. Then, asking a close relative, the dietary habits were ascertained.

	High Salt	Low Salt	Total
CVD	5	30	35
Non-CVD	2	23	25
Total	7	53	60

A calculation of the odds ratio yields:

$$OR = \frac{5 \times 23}{2 \times 30} = 1.92$$

Fisher's Exact Test

	D+	D-	Total
E+			n_1
E-			n_2
Total	m_1	m_2	N

If we **fix all of the margins** then any one cell of the table will allow the remaining cells to be filled. Note that *a* must be greater than 0, less than both n_1 and m_1, and an integer. Thus there are only a relatively few number of possible table configurations if either $n_1$or m_1 is small (with n_1, n_2, m_1, m_2 fixed).

Under the null hypothesis,

H_0 : OR = 1

we can use the hypergeometric distribution (a probability distribution for discrete rv's) to compute the probability of any given configuration. Since we have the distribution of a statistic (*a*) under the null, we can use this to compute p-values. You will *never* do this by hand ….

Fisher's Exact Test

Example: (Rosner, p. 370) Cardiovascular disease.

	High Salt	Low Salt	Total
CVD	5	30	35
Non-CVD	2	23	25
Total	7	53	60

Possible Tables:

0		35
		25
7	53	60

1		35
		25
7	53	60

2		35
		25
7	53	60

3		35
		25
7	53	60

4		35
		25
7	53	60

5		35
		25
7	53	60

6		35
		25
7	53	60

7		35
		25
7	53	60

Fisher Exact Test Using Stata

Fisher's exact test.

```
. cci 5 30 2 23,exact

                |  Exposed   Unexposed  |    Total   Proportion Exposed
----------------+-----------------------+-------------------------------
          Cases |          5          30 |         35             0.1429
       Controls |          2          23 |         25             0.0800
----------------+-----------------------+-------------------------------
          Total |          7          53 |         60             0.1167
                |                       |
                |      Point estimate   |    [95% Conf. Interval]
                |-----------------------+-------------------------------
     Odds ratio |          1.916667     |     .2789585    21.62382  (exact)
Attr. frac. ex. |          .4782609     |    -2.584763    .9537547  (exact)
Attr. frac. pop |          .068323      |
                +-----------------------+

                                  1-sided Fisher's exact P = 0.3747
                                  2-sided Fisher's exact P = 0.6882
```

Fisher Exact Test Using Stata

The usual chi-squared test, for comparison.

```
. cci 5 30 2 23

                                                    Proportion
                |  Exposed   Unexposed  |  Total     Exposed
----------------+-----------------------+-----------------------
          Cases |        5          30  |     35      0.1429
       Controls |        2          23  |     25      0.0800
----------------+-----------------------+-----------------------
          Total |        7          53  |     60      0.1167
                |                       |
                |      Point estimate   |    [95% Conf. Interval]
                |-----------------------+-----------------------
      Odds ratio |        1.916667      |    .2789585   21.62382  (exact)
 Attr. frac. ex. |        .4782609      |   -2.584763   .9537547  (exact)
 Attr. frac. pop |        .068323       |
                +-----------------------+

                                  chi2(1) =    0.56  Pr>chi2 = 0.4546
```


False Discovery Rate

For some studies, answering the scientific question of interest may require testing hundred, thousands, or millions of hypotheses. This is especially true of genetics.

E.g. Hedenfalk et al (2001) screened 3226 genes using microarrays to find differential expression between BRCA-1 and BRCA-2 mutation positive tumors.

Issue: If a traditional hypothesis testing approach is taken and we conduct 3226 tests at the 0.05 level, then we expect (up to) 161 false positive findings. Unfortunately, they are not labeled as such!

Traditional Solution (Bonferroni correction): If we conduct each test at an $\alpha = .05/3226 = .000015$ level then the probability of 1 or more false positive findings will be ~0.05. But, … with such a stringent α level we are likely to miss many true positive results.

New Solution: Don't try to eliminate false positives … control them

False Discovery Rate

	Reject null	Fail to reject	
Null true	F	m_0-F	m_0
Alternative true	T	m_1-T	m_1
	S	m-S	m

- false positive rate = F/ m_0
- false discovery rate = F/S

Idea: Control the false discovery rate (q-value) instead of the false positive rate (p-value)

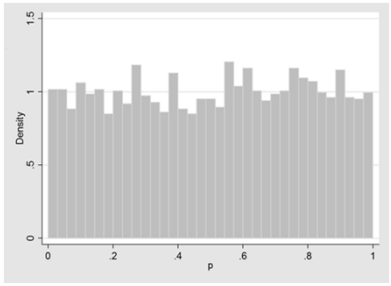
False Discovery Rate

E.g. Hedenfalk data

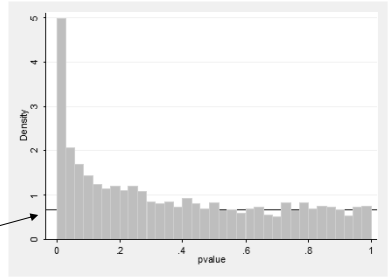
- Order the 3170 p-values (56 genes were excluded from this analysis): p_i , i = 1 … 3170
- Pick a p-value cutoff, say α; reject H_o for all $p_i < \alpha$.
- From the previous table we know S = #{$p_i < \alpha$}
- Also, F = α * m_0
- FDR = F/S … I know S, I know α, what is m_0?

False Discovery Rate

Distribution of 3170 p-values when all null hypotheses are true



Distribution of 3170 p-values from Hedenfalk et al. Height of the line gives estimated proportion of true null hyptheses.



$$m_0(\lambda) = \frac{\#\{p_i > \lambda; i = 1...m\}}{(1-\lambda)}$$

False Discovery Rate

- $q(\alpha) = FDR(\alpha;\lambda) = \alpha * m_0(\lambda)/\#\{p_i < \alpha\}$
  (technically $q(\alpha) = \min_{t \geq \alpha} FDR(t)$)
- Program QVALUE (http://genomine.org/qvalue/)
- Eg. Hedenfalk et al. ($m_0(.5) = 2143$)

No. differentially expressed

q	α	$\#\{p_i < \alpha\}$	expected false pos
.01	.0000126	5	0
.05	.00373	160	8
.10	.0148	317	32

- Using traditional methods Hedenfalk et al. concluded 9-11 genes were differentially expressed.