Chapter 5: The Role of Probability

Dr. Daniel B. Rowe
Professor of Computational Statistics
Department of Mathematical and Statistical Sciences
Marquette University

Copyright D.B. Rowe

Probability

Probabilities are numbers that reflect the likelihood that a particular event occurs.

Statistical inference involves making generalizations or inferences about unknown population parameters based on sample statistics.

Parameters: Summary measures computed on populations. i.e. μ, σ^2

Statistics: Numerical summary measures computed on samples. i.e. $\bar{X}$, s^2

5.1 Sampling

Sampling Frame: A complete list or enumeration of the population.

Simple Random Sampling: A set of numbers is selected at random to determine the individuals to be included.

Systematic Sampling: Individuals selected at regular interval N/n.
N is population size, n is desired sample size.
i.e. very third or fifth selected. Might not be representative.

5.1 Sampling

Stratified Sampling: Split the population into nonoverlapping groups or strata then sample within each stratum.
Instead of randomly from entire US population, sample proportionately from each state.

Convenience Sampling: Select individuals by any convenient contact.
Select patients as they come in, not from all patients.

5.2 Basic Concepts

Probability is a number that reflects the likelihood that a particular event Will occur. Probabilities range from 0 to 1.

$$P(characteristic) = \frac{Number \quad of \quad persons \quad with \quad characteristic}{Total \quad number \quad of \quad persons \quad in \quad the \quad population \quad (N)}$$

	Age (years)						
	5	6	7	8	9	10	Total
Boys	432	379	501	410	420	418	2560
Girls	408	513	412	436	461	500	2730
Total	840	892	913	846	881	918	5290

$$P(boy) = \frac{2560}{5290} = 0.484$$

5.3 Conditional Probability

Sometimes it is of interest to focus on a particular subset of the population.

What is the probability of selecting a 9-year-old girl from the subpopulation of girls?

	Age (years)						
	5	6	7	8	9	10	Total
Boys	432	379	501	410	420	418	2560
Girls	408	513	412	436	461	500	2730
Total	840	892	913	846	881	918	5290

$$P(9-year-old \mid girls) = \frac{461}{2730} = 0.169$$

16.9% of girls are 9-years old.

5.3 Conditional Probability

Screening tests are often used in clinical practice. Results changes probs.

What is the probability of a male having prostate cancer?

	Biopsy Results		
	Prostate Cancer	No Prostate Cancer	Total
Total	28	92	120

Abbreviation: PSA, prostate-specific antigen.

$$P(\textit{prostate cancer}) = \frac{28}{120} = 0.233$$

5.3 Conditional Probability

Screening tests are often used in clinical practice. Results changes probs.

What is the probability of a male having prostate cancer?

	Biopsy Results		
PSA Level	Prostate Cancer	No Prostate Cancer	Total
Low	3	61	64
Slightly to moderately elevated	13	28	41
Highly elevated	12	3	15
Total	28	92	120

Abbreviation: PSA, prostate-specific antigen.

$$P(prostate\ cancer) = \frac{28}{120} = 0.233$$

$$P(prostate\ cancer \mid low\ PSA) = \frac{3}{64} = 0.047$$

PSA = prostate-specific antigen

5.3 Conditional Probability

Screening tests are often used in clinical practice. Results changes probs.

What is the probability of a male having prostate cancer?

PSA Level	Biopsy Results: Prostate Cancer	Biopsy Results: No Prostate Cancer	Total
Low	3	61	64
Slightly to moderately elevated	13	28	41
Highly elevated	12	3	15
Total	28	92	120

Abbreviation: PSA, prostate-specific antigen.

$$P(prostate\ cancer) = \frac{28}{120} = 0.233$$

$$P(prostate\ cancer \mid low\ PSA) = \frac{3}{64} = 0.047$$

$$P(prostate\ cancer \mid slight\ to\ moderate\ PSA) = \frac{13}{41} = 0.317$$

PSA = prostate-specific antigen

5.3 Conditional Probability

Screening tests are often used in clinical practice. Results changes probs.

What is the probability of a male having prostate cancer?

PSA Level	Biopsy Results: Prostate Cancer	Biopsy Results: No Prostate Cancer	Total
Low	3	61	64
Slightly to moderately elevated	13	28	41
Highly elevated	12	3	15
Total	28	92	120

Abbreviation: PSA, prostate-specific antigen.

$$P(\text{prostate cancer}) = \frac{28}{120} = 0.233$$

$$P(\text{prostate cancer} \mid \text{low PSA}) = \frac{3}{64} = 0.047$$

$$P(\text{prostate cancer} \mid \text{slight to moderate PSA}) = \frac{13}{41} = 0.317$$

$$P(\text{prostate cancer} \mid \text{highly elevated PSA}) = \frac{12}{15} = 0.80$$

PSA = prostate-specific antigen

5.3 Conditional Probability

Sensitivity is also called the true positive fraction.

Specificity is also called the true negative fraction.

	Disease present	Disease Free	Total
Screen positive	a	b	$a+b$
Screen negative	c	d	$c+d$
Total	$a+c$	$b+d$	N

$$Sensitivity = True\ Positive\ Fraction = P(screen\ positive \mid disease) = \frac{a}{a+c}$$

$$Specificity = True\ Negative\ Fraction = P(screen\ negative \mid disease\ free) = \frac{d}{b+d}$$

$$False\ Positive\ Fraction = P(screen\ positive \mid disease\ free) = \frac{b}{b+d}$$

$$False\ Negative\ Fraction = P(screen\ negative \mid disease) = \frac{c}{a+c}$$

5.3 Conditional Probability

Consider the N=4810 pregnancies with blood screen & amniocentesis for likelihood of Down Syndrome.

	Affected Fetus	Unaffected Fetus	Total
Positive	9	351	360
Negative	1	4449	4450
Total	10	4800	4810

$$Sensitivity = P(screen\ \ positive \,|\, affected\ \ \ fetus) = \frac{9}{10} = 0.900$$

$$Specificity = P(screen\ negative \,|\, unaffected\ \ \ fetus) = \frac{4449}{4800} = 0.927$$

$$FP\ Fraction = P(screen\ \ positive \,|\, unaffected\ \ fetus) = \frac{351}{4800} = 0.073$$

$$FN\ \ Fraction = P(screen\ negative \,|\, affected\ \ fetus) = \frac{1}{10} = 0.100$$

5.4 Independence

Two events are **independent** if the probability of one is not affected by the occurrence or nonoccurrence of the other.

	Biopsy Results		
Prostate Test Risk	Prostate Cancer	No Prostate Cancer	Total
Low	10	50	60
Moderate	6	30	36
High	4	20	24
Total	20	100	120

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

A=Low Risk

B= Prostate Cancer

$$P(A|B)=P(low\ risk\ |\ prostate\ cancer)=\frac{10}{20}=0.50$$

$$P(A)=P(low\ risk)=\frac{60}{120}=0.50$$

A and B are Independent

5.5 Bayes Theorem

Bayes Theorem is a probability rule to compute conditional probabilities.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Example: Patient exhibiting symptoms of rare disease.

$$P(disease\,|\,screen\ \ positive) = \frac{P(screen\ \ positive\,|\,disease)P(disease)}{P(screen\ \ positive)}$$

$$\left.\begin{array}{l} P(disease) = 0.002 \\ P(screen\ \ positive\,|\,disease) = 0.85 \\ P(screen\ \ positive) = 0.08 \end{array}\right] \longrightarrow P(disease\,|\,screen\ \ positive) = \frac{(0.85)(0.002)}{(0.08)} = 0.021$$

5.6 Probability Models – Binomial Distribution

$P(H)+P(T)=1$

Let's assume we are flipping a coin twice.

H=Head on flip, T=Tail on flip

The probability of heads on any given flip is $p = P(H)$.

The probability of tails (not heads) on any given flip is $q = (1-p)$.

Independent events

Independent events

Then $P(HT)=P(H)P(T)$ $=p(1-p)$. Similarly $P(TH) = P(T)P(H)$ $= (1-p)p$.

Let x = # of heads in two flips of a coin.

consider both ways

$P(x=1) = P(HT)+P(TH)$

2 ways to get one H and one T
2 ways to get x=1 heads

$= p(1-p)+(1-p)p = 2p(1-p)$.

5.6 Probability Models – Binomial Distribution

An experiment with only two outcomes is called a Binomial experiment.

Call one outcome *Success* and the other *Failure*.

Each performance of experiment is called a trial and are independent.

$$P(x\ successes) = \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x}$$

Only for Binomial

$$\mu = np$$

$$\sigma^2 = np(1-p)$$

n = number of trials or times we repeat the experiment.

x = the number of successes out of n trials.

p = the probability of success on an individual trial.

$$\binom{n}{x} = \frac{n!}{x!(n-x)!}$$

5.6 Probability Models – Binomial Distribution

Example: Medication effectiveness.

$P(\textit{medication effective}) = p = 0.80$

What is the probability that it works on x=7 out of n=10?

$$P(7\ \ successes) = \frac{10!}{7!(10-7)!} 0.80^7 (1-0.80)^{10-7}$$

$$P(7\ \ successes) = \frac{10 \cdot 9 \cdot 8 \cdot \cancel{7!}}{\cancel{7!} 3 \cdot 2 \cdot 1} 0.80^7 0.20^3$$

$$P(7\ \ successes) = 120(0.2097)(0.008)$$

$$P(7\ \ successes) = 0.2013$$

$$P(x\ successes) = \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x}$$

n = number of trials or times we repeat the experiment.
x = the number of successes out of n trials.
p = the probability of success on an individual trial.

Questions?

Homework 5 Part I

Read Chapter 5.

Problems # 1*, 4

* What is the standard deviation σ of hyperlipidema?