Exam 1 Review

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

3.1 Prevalence

Prevalence refers to the proportion of participants with disease at a particular point in time.

An estimate of the prevalence of disease at baseline is

$$\text{Point Prevalence} = \frac{\text{Number of persons with disease}}{\text{Number of persons examined at baseline}}$$

3.1 Prevalence

Example 3.1 Computing Prevalence of Cardiovascular Disease (CVD)

TABLE 3–1 Men and Women with Diagnosed CVD

	Free of CVD	History of CVD	Total
Men	1548	244	1792
Women	1872	135	2007
Total	3420	379	3799

$$\text{Prevalence} = \frac{\text{\# with disease}}{\text{\# examined at baseline}}$$

Prevalence of CVD = 379/3799 = 0.0998 → 9.98%

Prevalence of CVD in Men = 244/1792 = 0.1362 → 13.62%

Prevalence of CVD in Women = 135/2007 = 0.0673 → 6.73%

3.2 Incidence

Incidence reflects the likelihood of developing disease among a group of participants free of the disease who are at risk of developing the disease over a specified observation period.

$$\text{Cumulative Incidence} = \frac{\text{Number of persons who develop disease during a specified period}}{\text{Number of persons at risk at baseline}}$$

$$\text{Incidence Rate} = \frac{\text{Number of persons who develop disease during a specified period}}{\text{Sum of the lengths of time during which persons are disease-free}}$$

3.2 Incidence

Cardiovascular Disease

Incidence of CVD?

Incidence Rate of CVD

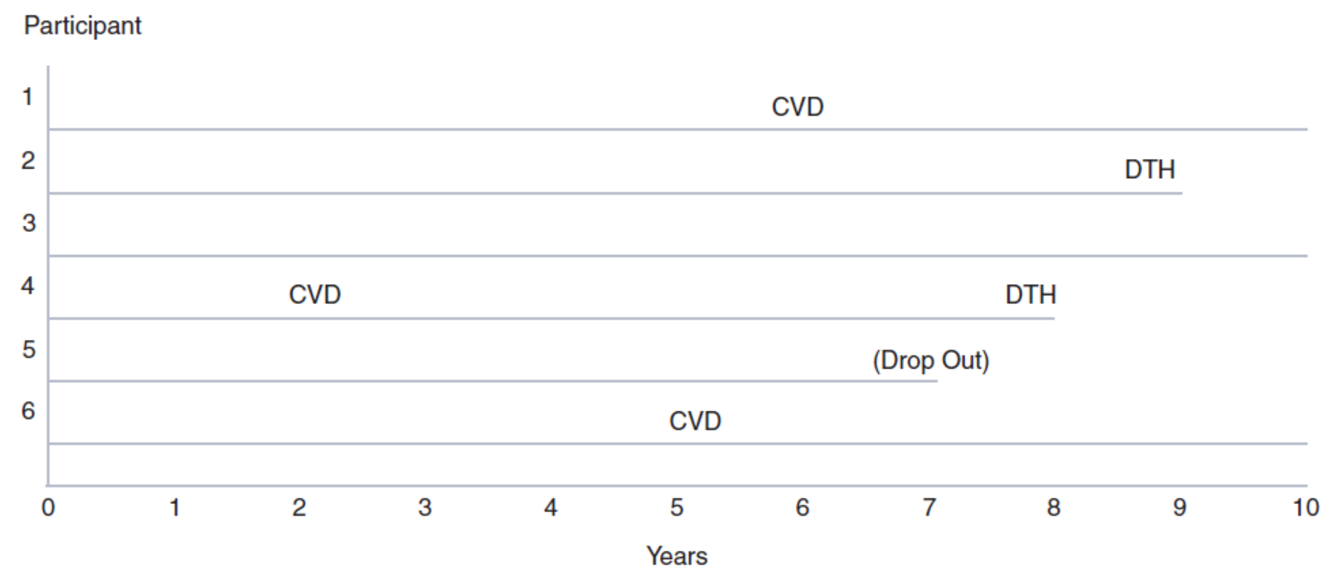
IR= 3/(6+9+10+2+7+5)

IR= 3/39

IR= 0.0769

7.7 per 100 person-years

FIGURE 3–1 Course of Follow-up for 6 Participants



3.4 Comparing Extent of Disease Between Groups

Cardiovascular Disease

Risk Difference of prevalent CVD in smokers versus nonsmokers

$$RD = \text{Prevalence}_{\text{smokers}} - \text{Prevalence}_{\text{nonsmokers}}$$

TABLE 3–2 Smoking and Diagnosed CVD

	Free of CVD	History of CVD	Total
Nonsmoker	2757	298	3055
Current smoker	663	81	744
Total	3420	379	3799

$$\text{Prevalence} = \frac{\#\text{ with disease}}{\#\text{ examined at baseline}}$$

RD= 81/744 – 298/3055 = 0.1089 – 0.0975 = 0.0114

3.4 Comparing Extent of Disease Between Groups

Relative Risk (RR) of CVD in smokers versus nonsmokers

$$RR = \frac{\text{Prevalence}_{\text{smokers}}}{\text{Prevalence}_{\text{nonsmokers}}} = \frac{81/744}{298/3055} = \frac{0.1089}{0.0975} = 1.12$$

TABLE 3–2 Smoking and Diagnosed CVD

	Free of CVD	History of CVD	Total
Nonsmoker	2757	298	3055
Current smoker	663	81	744
Total	3420	379	3799

$$\text{Prevalence} = \frac{\text{\# with disease}}{\text{\# examined at baseline}}$$

3.4 Comparing Extent of Disease Between Groups

Odds Ratio of CVD in hypertensives vs. non-hypertensives.

$$OR = \frac{181/840 \big/ (1-181/840)}{188/2942 \big/ (1-188/2942)} = \frac{0.275 \big/ 0.725}{0.068 \big/ 0.932} = 4.04$$

TABLE 3–5 Prevalent Hypertension and Prevalent CVD

	No CVD	CVD	Total
No hypertension	2754	188	2942
Hypertension	659	181	840
Total	3413	369	3782

$$\text{Prevalence} = \frac{\#\text{ with disease}}{\#\text{ examined at baseline}}$$

$$OR = \frac{\text{Prevalence}_{\text{exposed}} \big/ (1-\text{Prevalence}_{\text{exposed}})}{\text{Prevalence}_{\text{unexposed}} \big/ (1-\text{Prevalence}_{\text{unexposed}})}$$

Data

The **population** is the collection of all individuals about whom we wish to make generalizations.
The **sample** is a subset of individuals from the population.

Dichotomous variables have only two possible responses. Yes/No
Ordinal variables have more than two possible ordered responses.
Categorical variables sometimes called nominal variables are similar to ordinal variables except that the responses are unordered.

Data

Continuous variables take on an unlimited number of responses between defined minimum and maximum values.

Statistics: Numerical summary measures computed on samples.

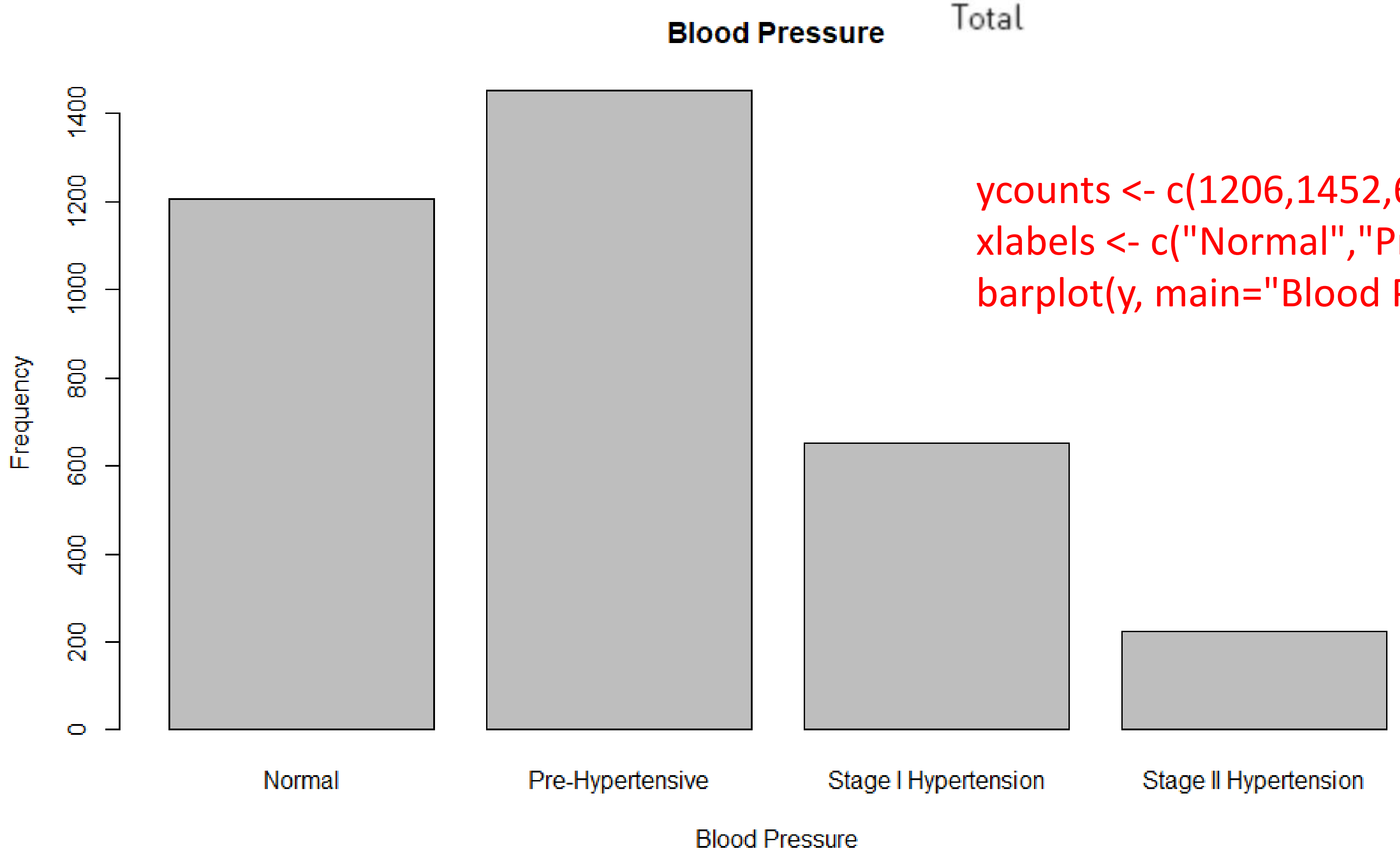
Parameters: Summary measures computed on populations.

4.2 Ordinal and Categorical Variables

Example:

	Frequency	Relative Frequency (%)	Cumulative Frequency	Cumulative Relative Frequency (%)
Normal	1206	34.1	1206	34.1
Prehypertension	1452	41.1	2658	75.2
Stage I hypertension	653	18.5	3311	93.7
Stage II hypertension	222	6.3	3533	100.0
Total	3533	100.0		

Description




```
ycounts <- c(1206,1452,653,222)
xlabels <- c("Normal","Pre-Hypertensive","Stage I Hypertension", "Stage II Hypertension")
barplot(y, main="Blood Pressure",xlab="Blood Pressure", ylab="Frequency",names.arg=xlabels)
```

4.3 Continuous Variables

Example 1: Small Numbers

Data values: 1,2,2,3,4

Sample Mean

Notation for sum x's

$$\bar{X} = \frac{\sum X}{n} = \frac{12}{5} = 2.4$$

$$\sum X = 1+2+2+3+4=12$$

Notation for sum x's

```
x <- c(1,2,2,3,4)
sum(x)
mean(x)
```

4.3 Continuous Variables

Example 1: Small Numbers

Data values: 1,2,2,3,4

Sample Median

$$median = middle \quad value$$
$$median = 2$$

Order data from smallest to largest.
If the number of data values is odd, take the middle value as the median.
If the number of data values is even, take the average of the middle two.

Sample Mode

$$mode = most \quad frequent \quad value$$
$$mode = 2$$

Order data from smallest to largest.
Count how many time each value occurs. Take the one with the highest count.

4.3 Continuous Variables

X	$\bar{X}$	$X-\bar{X}$	$(X-\bar{X})^2$
1	2.4	-1.4	1.96
2	2.4	-0.4	0.16
2	2.4	-0.4	0.16
3	2.4	0.6	0.36
4	2.4	1.6	2.56
$\sum$ 12			5.20

Example 1: Small Numbers

Data values: 1,2,2,3,4

Sample Variance & Standard Deviation

$$s^2 = \frac{1}{n-1}\sum(X-\bar{X})^2$$

$$s^2 = \frac{1}{5-1}\left[(1-2.4)^2+(2-2.4)^2+(2-2.4)^2+(3-2.4)^2+(4-2.4)^2\right]$$

$$s^2 = \frac{5.2}{4} = 1.3$$

$$s = \sqrt{s^2} = \sqrt{1.3} = 1.14$$

Standard Deviation

4.3 Continuous Variables

Example 1: Small Numbers

Data values: 1,2,2,3,4

Sample Variance & Standard Deviation

	X	X^2
	1	1
	2	4
	3	9
	3	9
	4	16
$\sum$	12	34

$n=5$

$$s^2=\frac{1}{n-1}\left[\sum X^2-\frac{1}{n}\left(\sum X\right)^2\right]$$

$$s^2=\frac{1}{5-1}\left[34-\frac{12^2}{5}\right]$$

$$s^2=\frac{5.2}{4}=1.3$$

$$s=\sqrt{s^2}=\sqrt{1.3}=1.14$$

4.3 Continuous Variables

Example 1: Small Numbers

Data values: 1,2,3,4,5

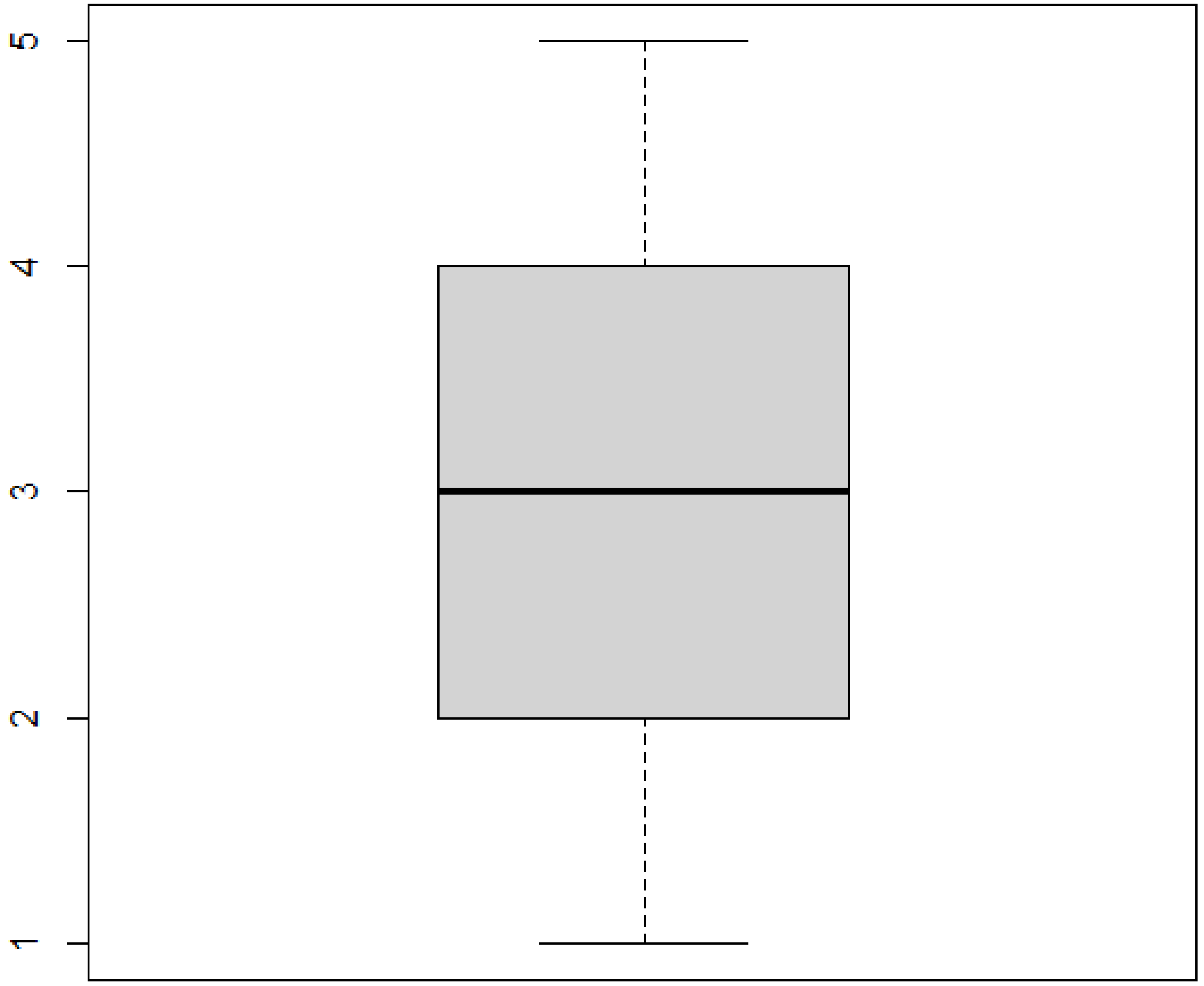
Box-Whisker Plot

5-number summary

1. L = minimum value
2. Q_1 = data value where 25% are smaller
3. Q_2 = median (where 50% are smaller)
4. Q_3 = data value where 75% are smaller
5. H = maximum value

$$IQR = Q_3 - Q_1$$

0%	25%	50%	75%	100%
1	2	3	4	5



Q_1 = median of lower half.
Q_3 = median of upper half

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

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

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

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

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

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

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.

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

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

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.

$P(medication\ effective)=p=0.80$

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?

Bring pencil, calculator, caffeinated beverage.

Will hand out exam and formula sheet.