Introduction to Medical Statistics 2026

Oxford University Clinical Research Unit
March 23-27, 2026

Ronald Geskus and the biostatistics crew
Oxford University Clinical Research Unit
Hospital for Tropical Diseases,
Ho Chi Minh City, Viet Nam

Part III

Statistical Analysis: Main Concepts and Principles; Binomial Distribution

Contents:

1. Types of study questions
2. Sampling variation
3. Binomial distribution
4. Testing hypotheses

Outline

Types of study questions

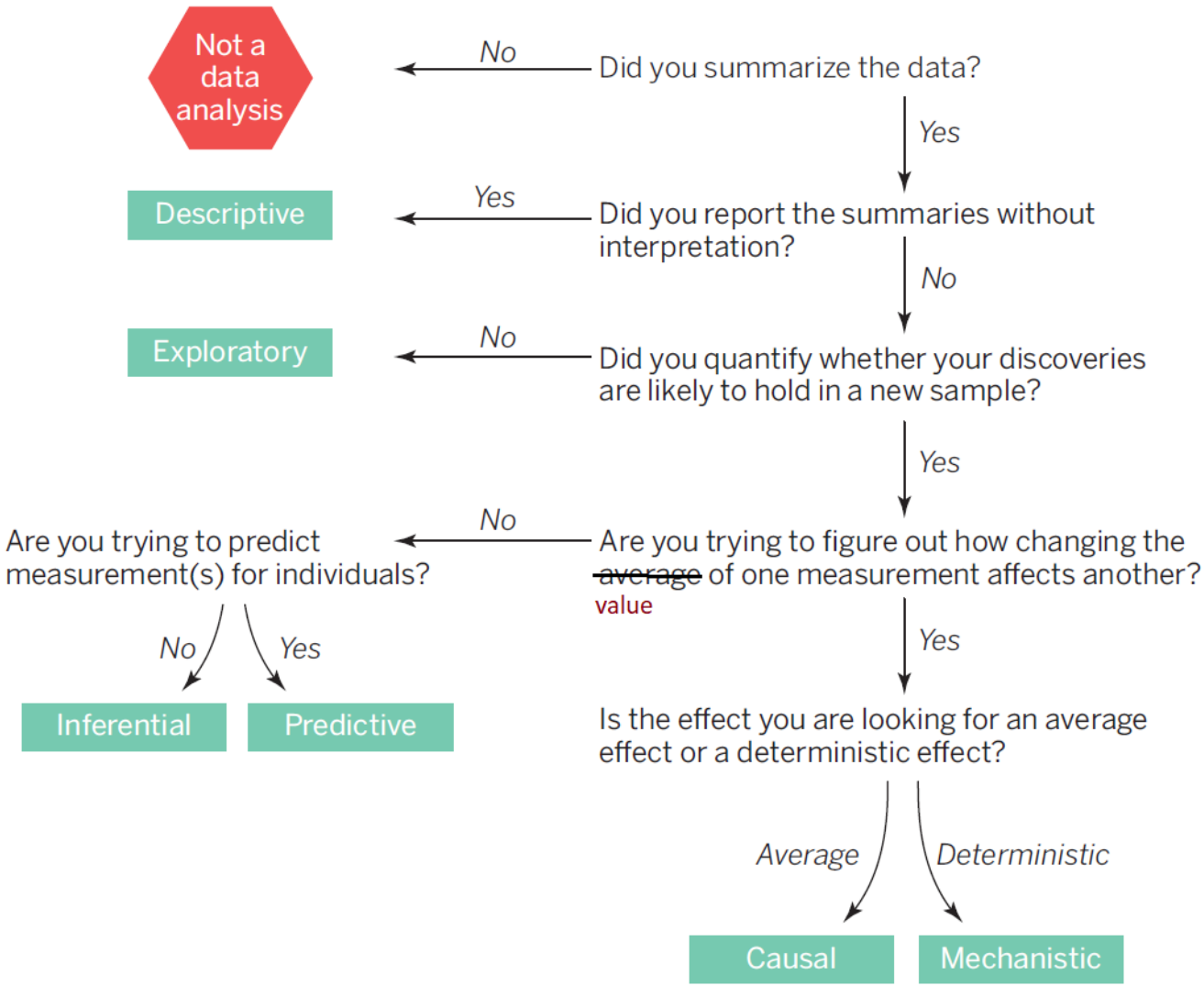
Sampling variation

Binary Variable; Estimating Proportions

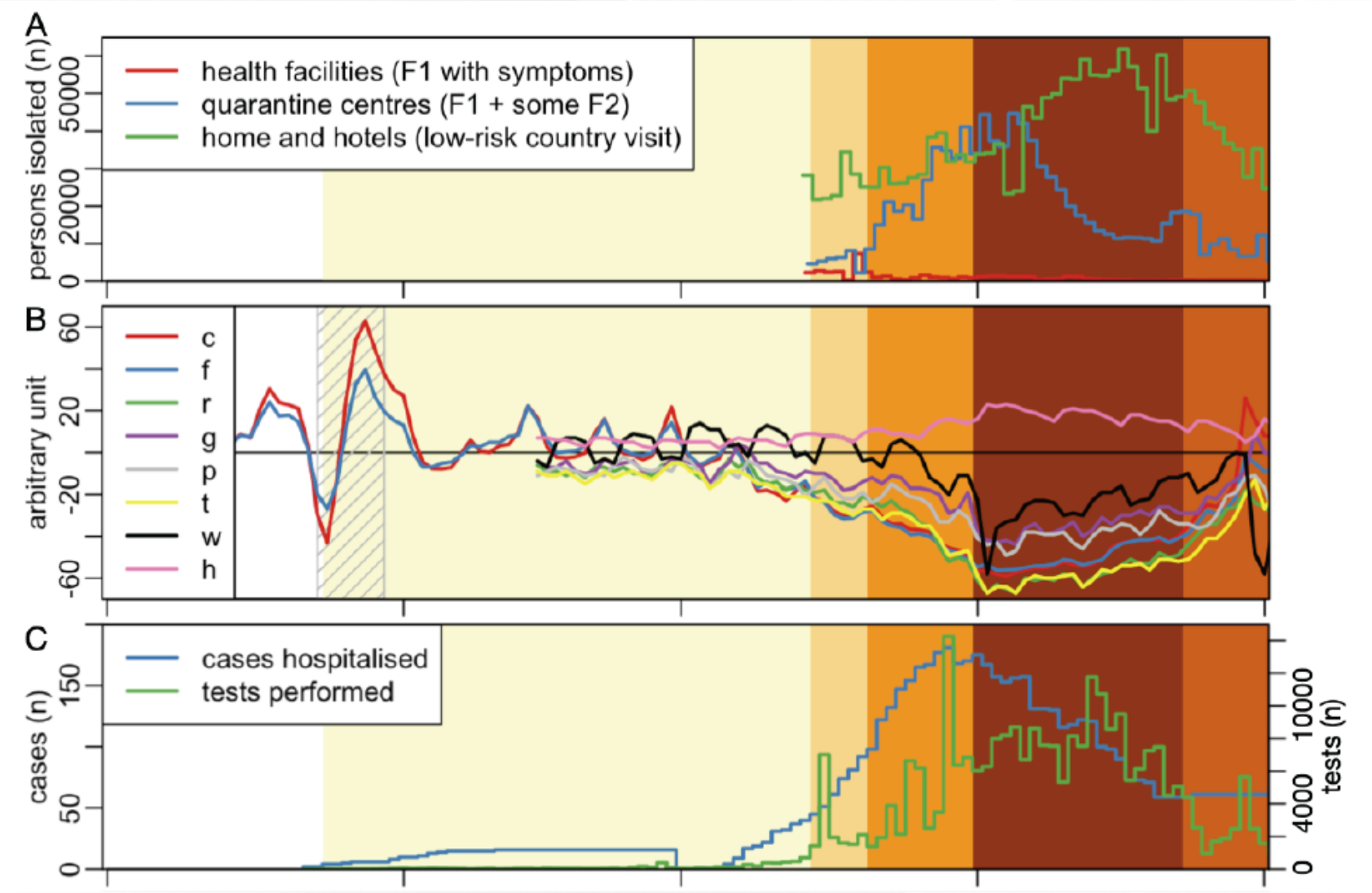
Testing a hypothesis

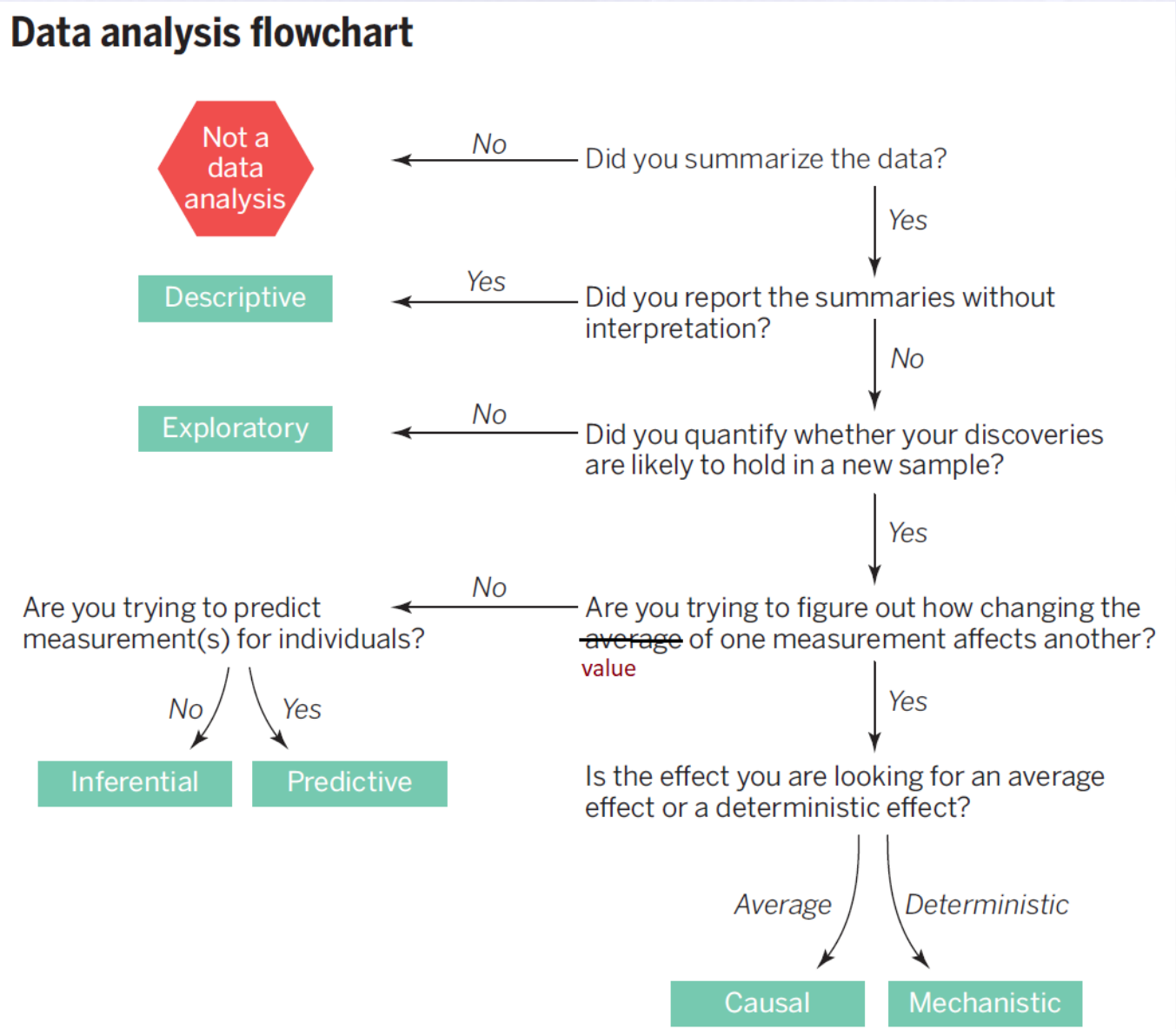
- Single proportion
- Compare proportions between two subgroups
- Categorical variables - more than 2 groups
- Alternative tests for specific settings

Data analysis flowchart



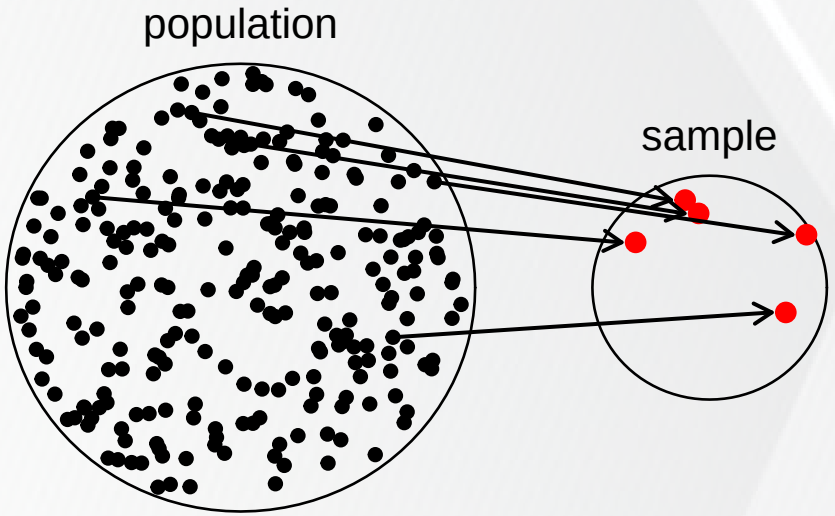
Exploratory: first 100 days of SARS-CoV-2 in Vietnam




Data analysis flowchart
Not a data analysis
No
Did you summarize the data?
Yes
Descriptive
Yes
Did you report the summaries without interpretation?
No
Exploratory
No
Did you quantify whether your discoveries are likely to hold in a new sample?
Yes
Are you trying to predict measurement(s) for individuals?
No
Are you trying to figure out how changing the ~~average~~ value of one measurement affects another?
No
Yes
Yes
Inferential
Predictive
Is the effect you are looking for an average effect or a deterministic effect?
Average
Deterministic
Causal
Mechanistic

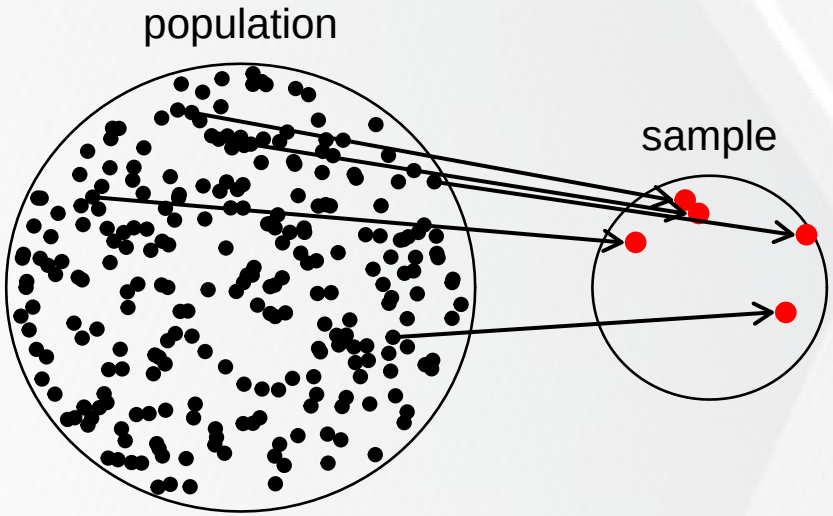
"Discovery likely to hold in a new sample"?

- Data (=sample) from (much) larger target population (often hypothetical, of "infinite size")



"Discovery likely to hold in a new sample"?

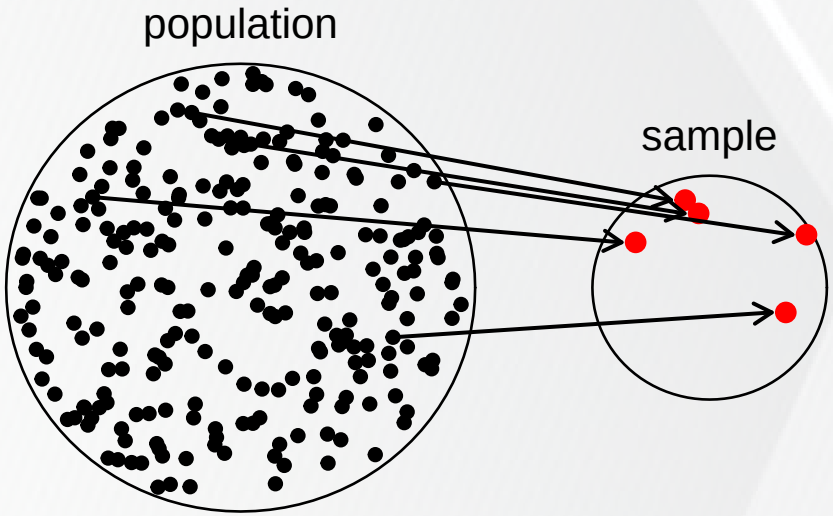
- Data (=sample) from (much) larger target population (often hypothetical, of "infinite size")



- Assumption: representative sample, reflects population. Result from sample can be generalized to population, i.e.
  - variable characteristics: sample $\approx$ population
  - relation between variables: in sample $\approx$ in population

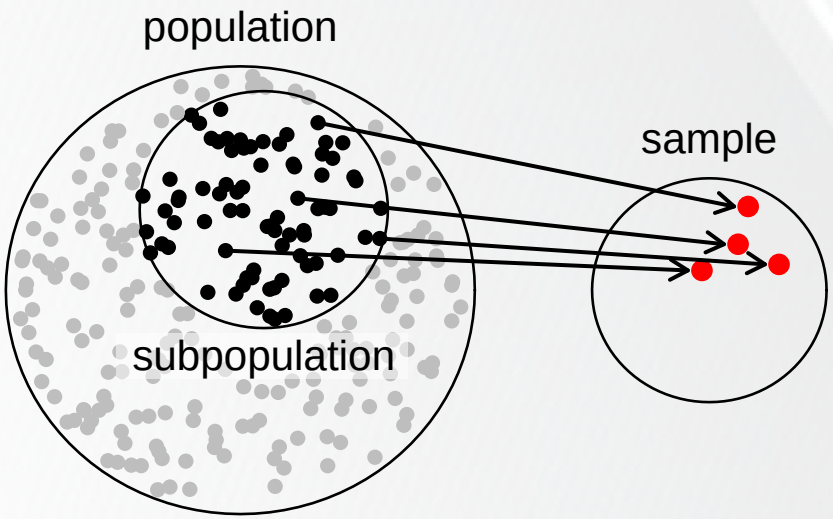
"Discovery likely to hold in a new sample"?

- Data (=sample) from (much) larger target population (often hypothetical, of "infinite size")

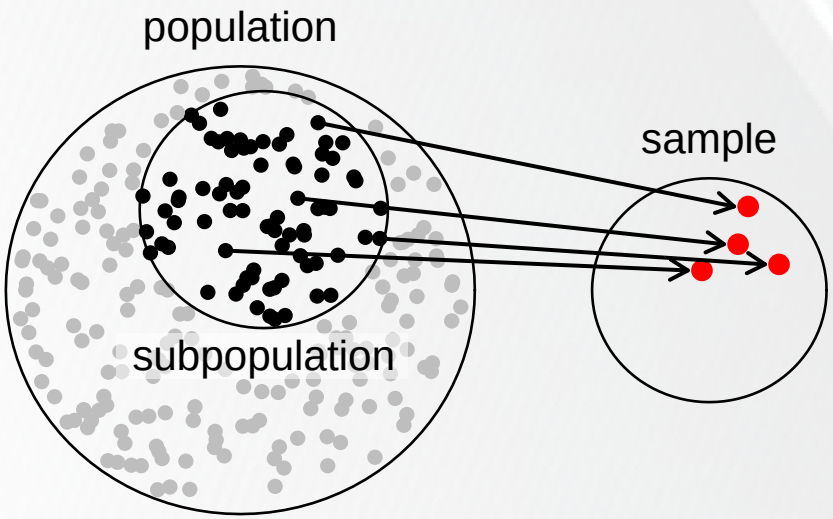


- Assumption: representative sample, reflects population. Result from sample can be generalized to population, i.e.
  - variable characteristics: sample $\approx$ population
  - relation between variables: in sample $\approx$ in population
- Similar result for a new representative sample, but some differences due to sampling variation

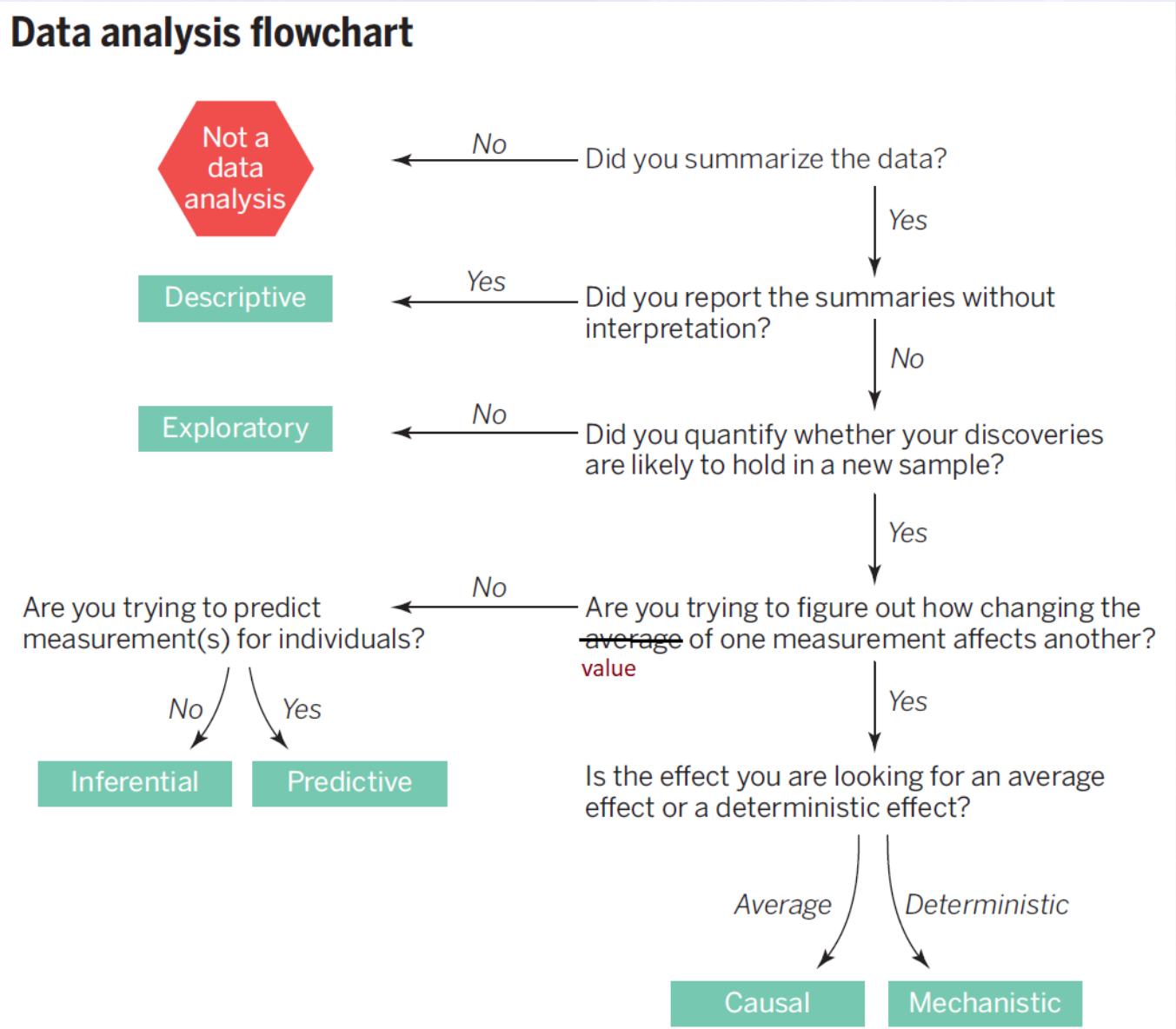
Generalize to which population?



Generalize to which population?



- Data: patients with dengue shock treated at HTD 2014-17
- Which population does the data represent?
  - all patients with dengue shock treated at HTD
  - all patients with dengue shock in Viet Nam
  - all patients with dengue shock, including patients who may present in the future and at other locations


Data analysis flowchart
Not a data analysis
No
Did you summarize the data?
Yes
Descriptive
Yes
Did you report the summaries without interpretation?
No
Exploratory
No
Did you quantify whether your discoveries are likely to hold in a new sample?
Yes
No
Are you trying to predict measurement(s) for individuals?
Are you trying to figure out how changing the ~~average~~ value of one measurement affects another?
No
Yes
Yes
Inferential
Predictive
Is the effect you are looking for an average effect or a deterministic effect?
Average
Deterministic
Causal
Mechanistic

Inferential, predictive or causal?

Relation between variable X and outcome Y

Inferential, predictive or causal?

Relation between variable X and outcome Y

Language: association, risk factor, predictor, effect, cause?

Inferential, predictive or causal?

Relation between variable X and outcome Y

Language: association, risk factor, predictor, effect, cause?

- *Smoking is associated with lung cancer*
- *Smoking is a risk factor for lung cancer*
- *Smoking is a predictor of lung cancer*
- *Smoking has an effect on lung cancer*
- *Smoking can cause lung cancer*

Inferential, predictive or causal?

Relation between variable X and outcome Y

Language: association, risk factor, predictor, effect, cause?

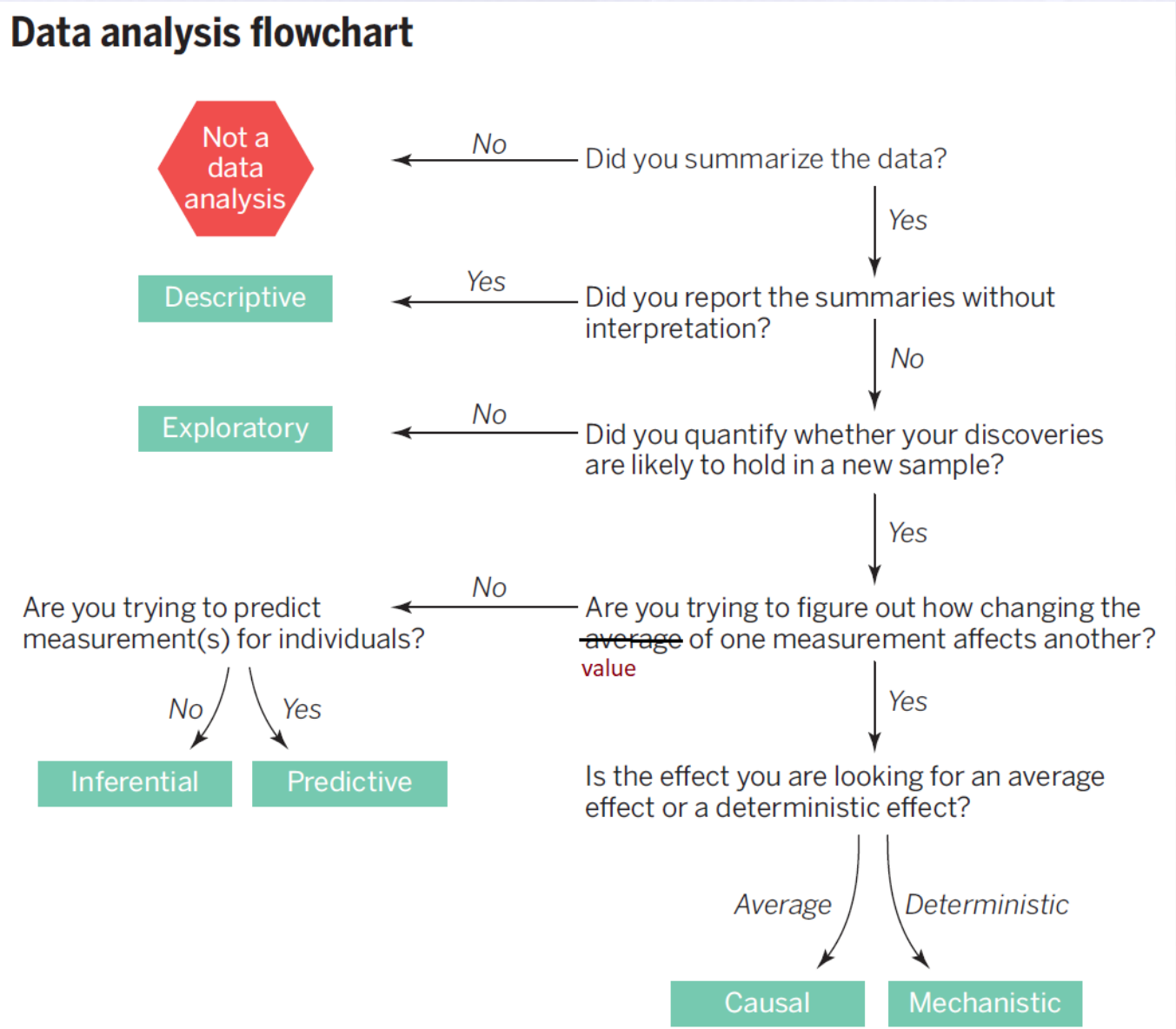
- *Carrying a lighter is associated with lung cancer*
- *Carrying a lighter is a risk factor for lung cancer*
- *Carrying a lighter is a predictor of lung cancer*
- *Carrying a lighter has an effect on lung cancer*
- *Carrying a lighter can cause lung cancer*

Association does not imply causation

- If x and y are associated, this may be because
  - x influences (or “causes”) y
  - y influences (or “causes”) x
  - both x and y are determined by one or more other variables
  - by chance
- See Chocolate Consumption, Cognitive Function, and Nobel Laureates and website on spurious correlations
- Causality is not a statistical concept. It requires use of background knowledge

THE FAMILY CIRCUS

"I wish they didn't turn on that seatbelt sign so much! Every time they do, it gets bumpy."


Data analysis flowchart
Not a data analysis
No
Did you summarize the data?
Yes
Descriptive
Yes
Did you report the summaries without interpretation?
No
Exploratory
No
Did you quantify whether your discoveries are likely to hold in a new sample?
Yes
No
Are you trying to predict measurement(s) for individuals?
Are you trying to figure out how changing the ~~average~~ value of one measurement affects another?
No
Yes
Yes
Inferential
Predictive
Is the effect you are looking for an average effect or a deterministic effect?
Average
Deterministic
Causal
Mechanistic

Example: tuberculous meningitis (TBM)

Relation between one or more variables and an outcome

Example: tuberculous meningitis (TBM)

Relation between one or more variables and an outcome

- Causal (well-defined hypothesis; “effect”)
  Does dexamethasone decrease mortality? (intervention)
  Role of Leukotriene A4 hydrolase (LTA4H) genotype in disease process (etiology)

Example: tuberculous meningitis (TBM)

Relation between one or more variables and an outcome

- Causal (well-defined hypothesis; “effect”)
  Does dexamethasone decrease mortality? (intervention)
  Role of Leukotriene A4 hydrolase (LTA4H) genotype in disease process (etiology)
- Inferential (risk factors; “association/correlation”)
  What are the risk factors for mortality? No formal causal structure specified

Example: tuberculous meningitis (TBM)

Relation between one or more variables and an outcome

- Causal (well-defined hypothesis; “effect”)
  Does dexamethasone decrease mortality? (intervention)
  Role of Leukotriene A4 hydrolase (LTA4H) genotype in disease process (etiology)
- Inferential (risk factors; “association/correlation”)
  What are the risk factors for mortality? No formal causal structure specified
- Predictive (personalized medicine; “prognostic/diagnostic value”)
  Prognostic: probability of dying within 12 months based on individual characteristics
  Diagnostic: probability to have TBM based on individual characteristics

Relationships between variables

- Without clear direction (correlation, association)
  - Risk factors for outcome (inferential)
  - Diagnosis based on test results and patient characteristics
- With direction $E \rightarrow D$
  - Risk factors for outcome (inferential)
  - Effect of exposure(s) E on outcome (causal)
  - Predict future outcome based on characteristics (prognostic)
- Role of variables in directional analysis
  - Dependent variable D, also called outcome, response, event
  - Independent variables E, also called covariates, covariables, predictors, risk factors, exposures, explanatory variables

Example: tuberculous meningitis (TBM)

Relation between one or more variables and an outcome

- Causal (well-defined hypothesis; “effect”)
  Does dexamethasone decrease mortality? (intervention)
  Role of Leukotriene A4 hydrolase (LTA4H) genotype in disease process (etiology)
- Inferential (risk factors; “association/correlation”)
  What are the risk factors for mortality? No formal causal structure specified
- Predictive (personalized medicine; “prognostic/diagnostic value”)
  Prognostic: probability of dying within 12 months based on individual characteristics
  Diagnostic: probability to have TBM based on individual characteristics

This course

- Focus on methods for inferential and causal questions
  - quantify relationship between variables and average value of outcome
  - p-value and confidence interval main concepts
  - Often via regression models
- Predictive questions: relationship between variables and outcome at individual level
  - predictive questions can be answered via regression models, but also via machine learning techniques
- Focus on frequentist estimation techniques (alternative: Bayesian statistics)

Outline

Types of study questions

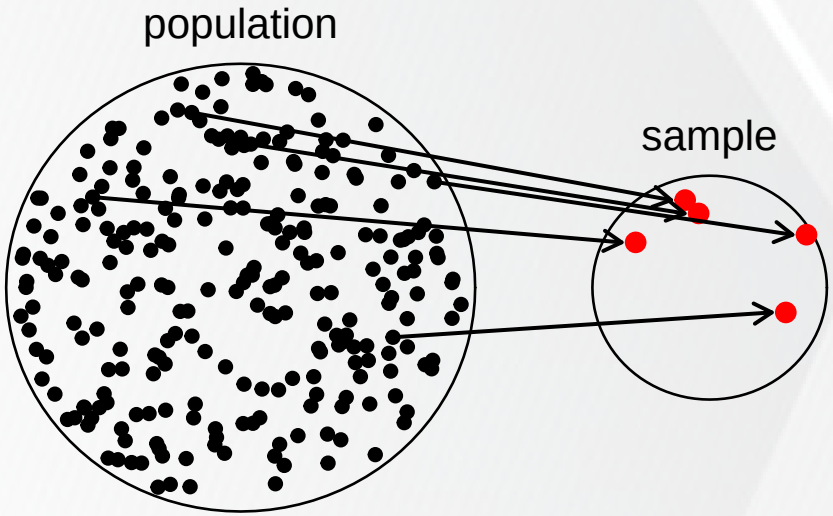
Sampling variation

Binary Variable; Estimating Proportions

Testing a hypothesis
- Single proportion
- Compare proportions between two subgroups
- Categorical variables - more than 2 groups
- Alternative tests for specific settings

"Discovery likely to hold in a new sample"?

- Data (=sample) from (much) larger target population (often hypothetical, of "infinite size")



- Assumption: representative sample, reflects population. Result from sample can be generalized to population, i.e.
  - variable characteristics: sample $\approx$ population
  - relation between variables: in sample $\approx$ in population
- Similar result for a new representative sample, but some differences due to sampling variation

Sample versus population: randomness

- Statistical science: develop methods to obtain estimate from sample that approximates true value in population

Sample versus population: randomness

- Statistical science: develop methods to obtain estimate from sample that approximates true value in population
- Sample data and estimate will always be
  - (slightly) different from data and estimate from new sample
  - (slightly) different from values in population
- Statistical science: develop methods to quantify uncertainty and variation in estimates

Sample versus population: randomness

- Statistical science: develop methods to obtain estimate from sample that approximates true value in population
- Sample data and estimate will always be
  - (slightly) different from data and estimate from new sample
  - (slightly) different from values in population
- Statistical science: develop methods to quantify uncertainty and variation in estimates
- Simplest setting: estimating mean value of a variable in the population

Sample versus population: terminology

- Parameter: characteristic of a population
  - Fixed value, but unknown
  - Notation: often Greek letter
    - π for event probability
      probability of dying after disease onset/hospitalisation
      (Covid-19, dengue, TBM,. . .)
    - μ for mean value of numeric variable
      mean viral load at time *t* after infection or disease onset

Sample versus population: terminology

- Parameter: characteristic of a population
  - Fixed value, but unknown
  - Notation: often Greek letter
    - π for event probability
      probability of dying after disease onset/hospitalisation
      (Covid-19, dengue, TBM,. . .)
    - μ for mean value of numeric variable
      mean viral load at time *t* after infection or disease onset
- Estimate: calculated from observations in sample
  - Assumed to represent (be close to) population value (parameter)
  - Value known in sample, but varies per sample
  - Notation: often with “hat”
    e.g. $\hat{p}$ for probability; $\hat{\bar{x}}$ for mean value;
    $\hat{\beta}$ for estimate of parameter in regression model

Sample versus population: terminology

- Parameter: characteristic of a population
  - Fixed value, but unknown
  - Notation: often Greek letter
    - π for event probability
      probability of dying after disease onset/hospitalisation
      (Covid-19, dengue, TBM,. . .)
    - μ for mean value of numeric variable
      mean viral load at time *t* after infection or disease onset
- Estimate: calculated from observations in sample
  - Assumed to represent (be close to) population value (parameter)
  - Value known in sample, but varies per sample
  - Notation: often with “hat”
    e.g. $\hat{p}$ for probability; $\hat{\bar{x}}$ for mean value;
    $\hat{\beta}$ for estimate of parameter in regression model
- More general term: statistic. Any function of the data that is used to infer on population characteristic
  In hypothesis testing we use a “test statistic”

Sampling variability

- $\hat{p}$ not exactly equal to π
  $\hat{\bar{x}}$ not exactly equal to μ
- new random sample from target population
  $\rightarrow$ new set of individuals
  $\rightarrow$ new value of estimate $\hat{p}$ or $\hat{\bar{x}}$

Sampling variability

- $\hat{p}$ not exactly equal to π
  $\hat{\overline{x}}$ not exactly equal to μ
- new random sample from target population
  $\rightarrow$ new set of individuals
  $\rightarrow$ new value of estimate $\hat{p}$ or $\hat{\overline{x}}$
- How accurate is our (single) estimate?
  sampling distribution: variation in estimate if we repeat the experiment (random sampling and computing estimate) many times

Sampling distribution

- Standard deviation describes variation/spread of any distribution
- Here: distribution of statistic if we repeatedly draw random samples from the population
- Standard deviation of statistic is called standard error

Sampling distribution

- Standard deviation describes variation/spread of any distribution
- Here: distribution of statistic if we repeatedly draw random samples from the population
- Standard deviation of statistic is called standard error
- If sample size is large then
  - statistic often follows a normal distribution (next class)
  - the amount of variation decreases and the estimate approaches population value

Outline

Types of study questions

Sampling variation

Binary Variable; Estimating Proportions

Testing a hypothesis

- Single proportion
- Compare proportions between two subgroups
- Categorical variables - more than 2 groups
- Alternative tests for specific settings

Binary variable

female/male; alive/death; no/yes; fail/success; 0/1

Examples

- negative NS1 ELISA test in patients with DENV infection
- development of severe dengue in patients with DENV infection

Setting

- Sample of n observations of a binary variable
- Count number of times X that "success" value occurs (negative NS1 ELISA, severe dengue)

The binomial distribution

Definition

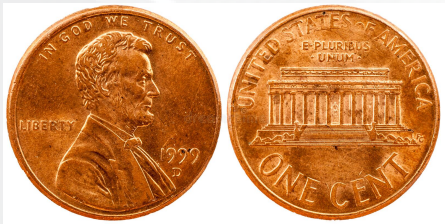
The variation in number of "successes" X in sample of n individuals. Notation: $X \sim B(n, p)$

- n: total number of observations/individuals
- p: probability of "success" per observation/individual
- X can be any whole number between 0 and n

The observations should meet these requirements:

1. The outcomes of all n observations are **independent** of each other.
2. All n observations have the **same probability of "success"**: p

Example



Flip a fair coin

- A coin is flipped 10 times. Outcome: head or tail
- The variable X is the number of heads among those 10 flips, our count of "successes"
- For each flip, the probability of success, "head", is known as $p = 0.5$
- Number X of heads has the binomial distribution $B(n = 10, p = 0.5)$

Binomial distribution: mean and variance of X

The mean μ and variance σ^2 of the binomial distribution for a count X are defined by the population success probability p via:

$$\mu = np$$

$$\text{variance} = np(1-p) \rightarrow$$

$$\text{sd} = \sqrt{np(1-p)}$$

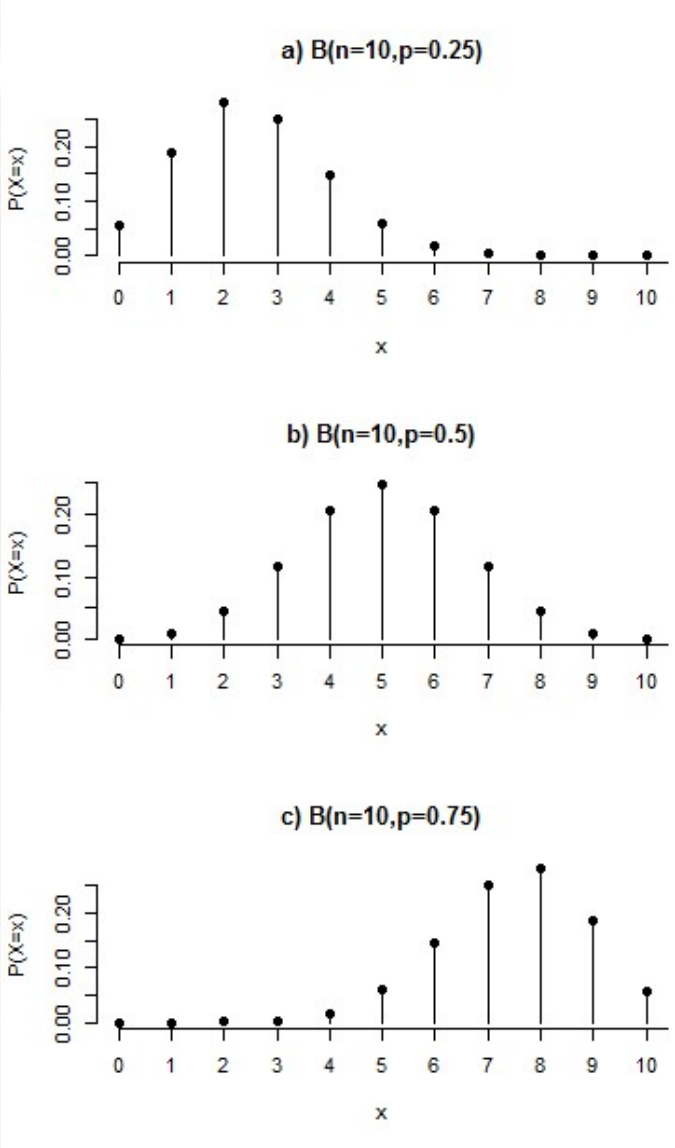
Binomial distribution: mean and variance of X

The mean μ and variance σ^2 of the binomial distribution for a count X are defined by the population success probability p via:

$$\mu = np$$

$$\text{variance} = np(1-p) \rightarrow$$

$$\text{sd} = \sqrt{np(1-p)}$$



Effect of changing p when n is fixed:

a $n = 10$, $p = 0.25$

b $n = 10$, $p = 0.5$

c $n = 10$, $p = 0.75$

Binomial distribution skewed when p different from 0.5 (especially if n is small).

Binomial distribution for $n = 3$

Examples

$n = 3$ patients, each can either have a success (S) or a failure (F)

Pt 1	Pt 2	Pt 3	Probability	Nr of successes
F	F	F	$(1-p)^3$	0
S	F	F	$p \cdot (1-p)^2$	1
F	S	F	$p \cdot (1-p)^2$	1
F	F	S	$p \cdot (1-p)^2$	1
S	S	F	$p^2 \cdot (1-p)$	2
S	F	S	$p^2 \cdot (1-p)$	2
F	S	S	$p^2 \cdot (1-p)$	2
S	S	S	p^3	3

$$X \sim B(n = 3, p) \rightarrow P(X = 2) = 3 \cdot p^2 \cdot (1-p)$$

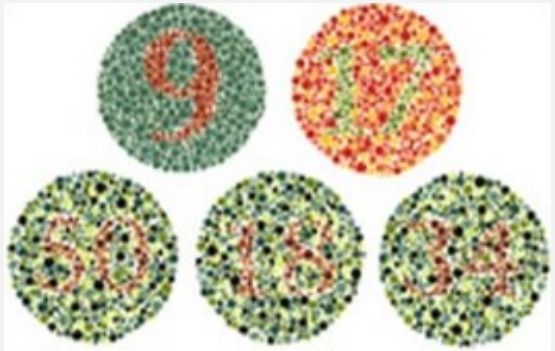
oucru

UNIVERSITY OF OXFORD

wellcome

Example: Color blindness

- Frequency of color blindness in Caucasian American male population is about 8%
- We take random sample of size 25 from this population



Probability that exactly 3 individuals in sample are color blind?

- Use R: `dbinom(3,size=25,prob=0.08)`
  ("d" in dbinom refers to "density/probability mass" function $P(X = k)$)
- Result: $P(X = 3) = 0.188$

Probability that 2 or fewer in the sample are color blind?

- Use R: `pbinom(2,size=25,prob=0.08)`
  ("p" in pbinom refers to the cumulative probability $P(X \leq k)$)
- Result: $P(X \leq 2) = 0.68$

Estimation of π

- If $X \sim B(n, p)$, then $\text{mean}(X) = np$ and $\text{sd}(X) = \sqrt{np(1-p)}$
- Estimation of population probability π: compute frequency $\hat{p} = X/n$
- Note: frequency is a mean. Let

$$X_i = \begin{cases} 1 & \text{if "success"} \\ 0 & \text{if "fail"} \end{cases}$$

then $\hat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i = X/n$, frequency of "success"

- $\text{sd}(\hat{p}) = \frac{1}{n}\text{sd}(X) = \frac{1}{n}\sqrt{n\pi(1-\pi)} = \sqrt{\frac{\pi(1-\pi)}{n}}$
  Decreases with increasing n

Outline

Types of study questions

Sampling variation

Binary Variable; Estimating Proportions

Testing a hypothesis

- Single proportion
- Compare proportions between two subgroups
- Categorical variables - more than 2 groups
- Alternative tests for specific settings

Outline

Types of study questions

Sampling variation

Binary Variable; Estimating Proportions

Testing a hypothesis

- Single proportion
- Compare proportions between two subgroups
- Categorical variables - more than 2 groups
- Alternative tests for specific settings

Example

- A new treatment for cancer
- Current treatments achieve a tumour response probability of 0.5
- In a clinical study of the new treatment with 20 patients, 15 showed a tumour response

Does this “prove” that the new treatment is better than the current treatments?

Formalisation of the problem

Probability model for the data

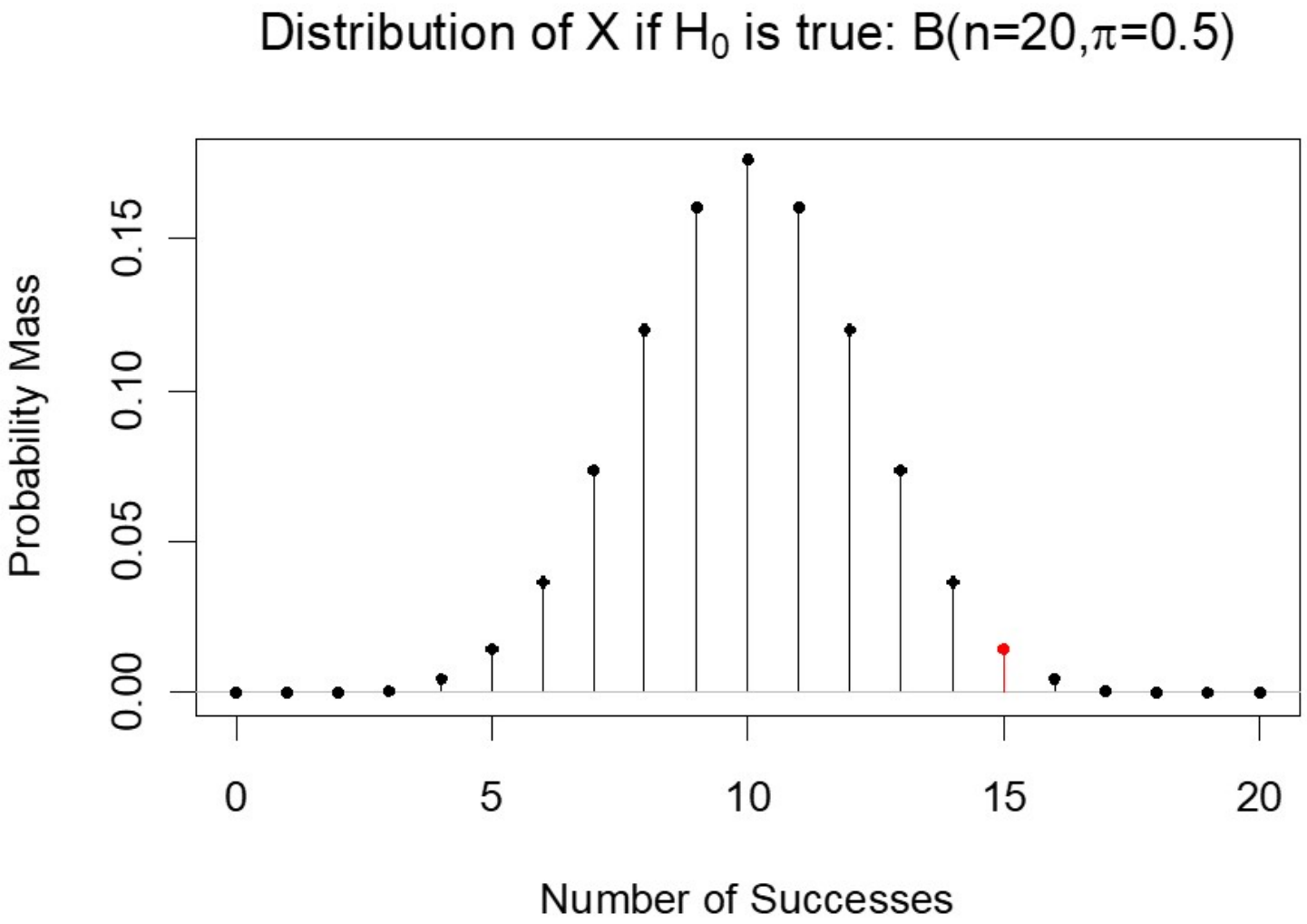
- The number of tumour responses $X \sim B(n = 20, \pi)$
  - Population success probability π
  - Variation based on $n = 20$ individuals in sample follows binomial distribution

Null hypothesis H_0

- $H_0 : \pi = 0.5$ (no effect)
  - $\hat{p} = 15/20$ successes occurred by chance. How likely is this?

Alternative hypothesis H_A

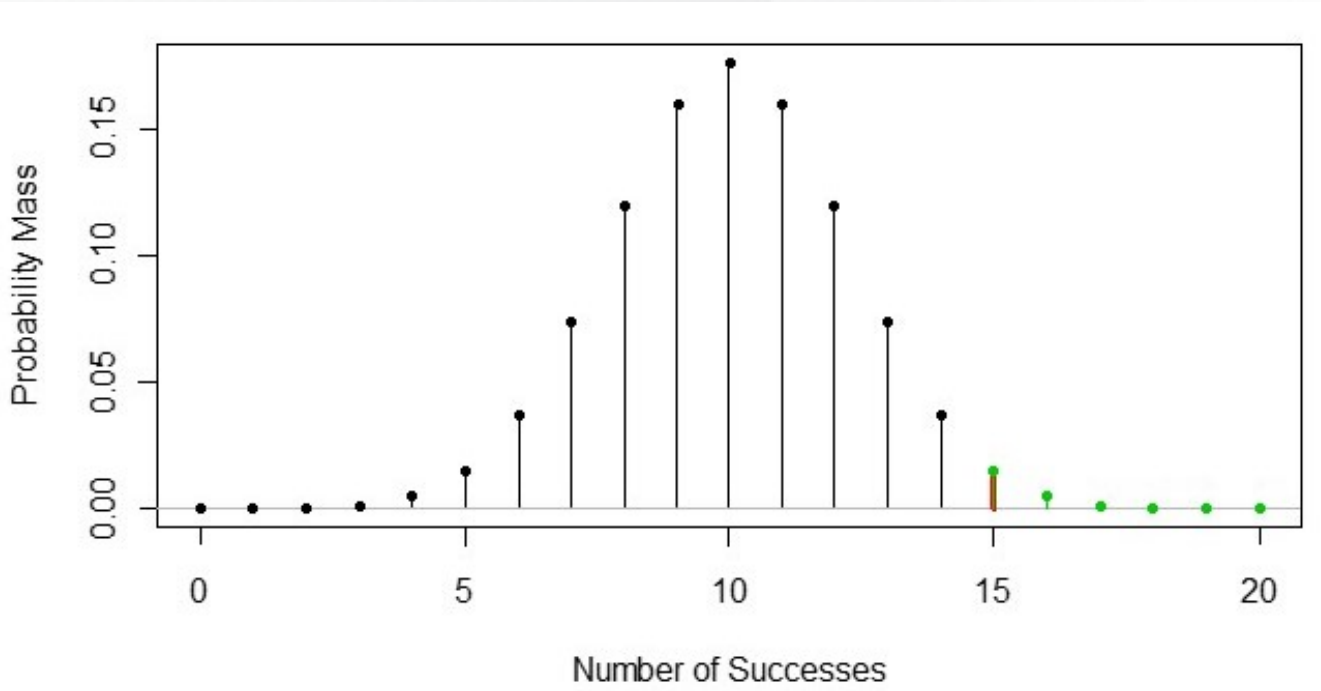
- H_A : True $\pi > 0.5$ (one-sided) - “better”
- H_A : True $\pi \neq 0.5$ (two-sided) - “different”


Distribution of X if H_0 is true: B(n=20,π=0.5)
Probability Mass
0.00
0.05
0.10
0.15
0
5
10
15
20
Number of Successes

P-value for one-sided alternative $H_A : \pi > 0.5$

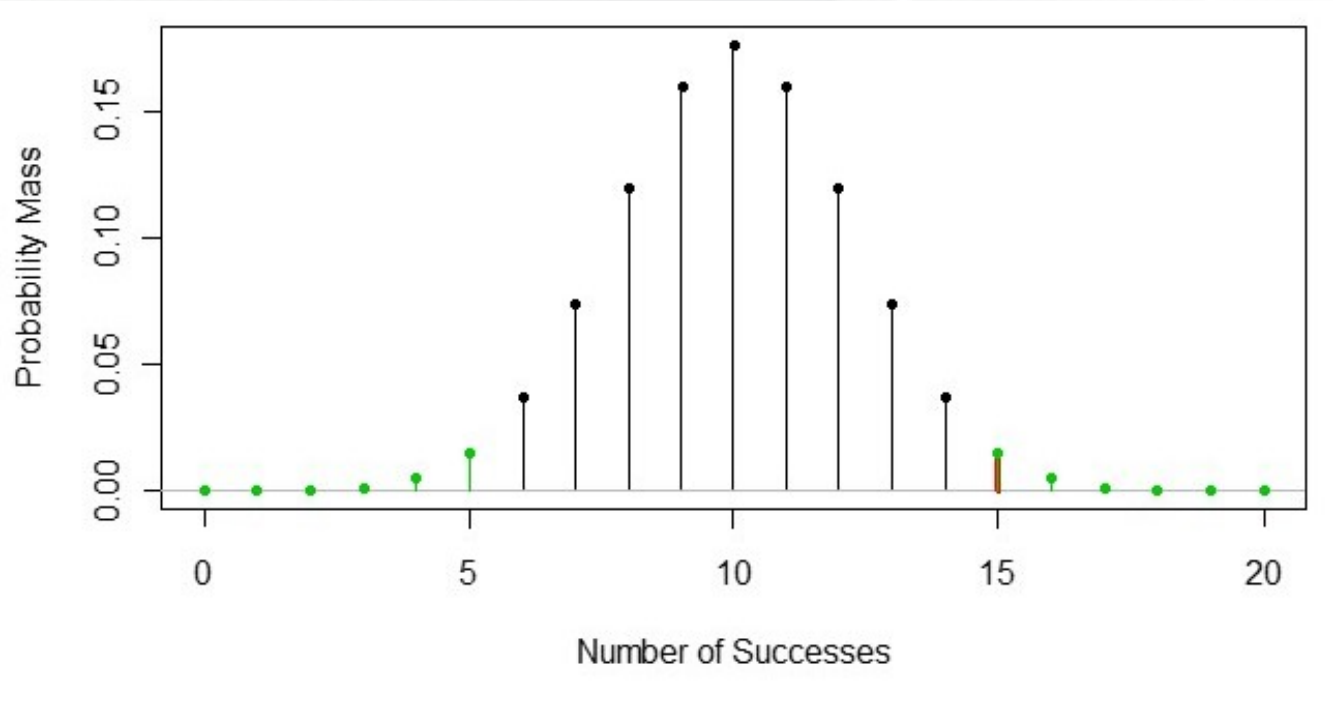
- How likely are 15/20 successes or more extreme ("**better**") if null hypothesis is true?
- If $\pi = 0.5$:

$$\begin{aligned} \text{p-value} &= P(X \geq 15) \\ &= P(X = 15) + P(X = 16) + ... + P(X = 20) \\ &= 0.014 + 0.004 + ... \\ &= 0.0207 \end{aligned}$$



P-value for two-sided alternative $H_A : \pi \neq 0.5$

- If the null hypothesis is true (which on average gives 10 successes), how likely to have deviation of 5 or more extreme
- "More extreme" : $X \leq 5$ or $X \geq 15$
- If $\pi = 0.5$: p-value = $P(X \leq 5) + P(X \geq 15) = 0.041$



P-value for the example with R

```
> prop.test(x=15,n=20,p=0.5,alternative="two.sided")

        1-sample proportions test with continuity correction

data:  15 out of 20, null probability 0.5
X-squared = 4.05, df = 1, p-value = 0.04417
alternative hypothesis: true p is not equal to 0.5
95 percent confidence interval:
 0.5058845 0.9040674
sample estimates:
   p
0.75
```

- For one-sided tests, use alternative="greater" or alternative="less"
- Reject null hypothesis if p-value is below a certain significance level α

One- and two-sided tests

- With symmetric distribution:
  - p-value (two-sided test) = 2 $\times$ p-value(one-sided test)

One- and two-sided tests

- With symmetric distribution:
  - p-value (two-sided test) = 2 $\times$ p-value(one-sided test)
- Two-sided alternatives are generally preferred
  - We cannot a priori exclude harm of a new treatment
  - One-sided p-value always smaller than 0.5

One- and two-sided tests

- With symmetric distribution:
  - p-value (two-sided test) = 2 $\times$ p-value(one-sided test)
- Two-sided alternatives are generally preferred
  - We cannot a priori exclude harm of a new treatment
  - One-sided p-value always smaller than 0.5
- Hence
  - Either two-sided tests at significance level α
  - Or one-sided tests at significance level $\alpha/2$
  - Often, $\alpha = 0.05$ is chosen

Hypothesis testing, basic structure

Effect exposure **E** on disease **D**

$$\mathbf{E} \longrightarrow \mathbf{D} \quad \text{in population}$$

Hypothesis testing, basic structure

Effect exposure **E** on disease **D**

$$\mathbf{E} \longrightarrow \mathbf{D} \quad \text{in population}$$

1. Null hypothesis H_0: no effect (often "$\beta = 0$")
2. Alternative hypothesis H_1: there is an effect (often "$\beta \neq 0$")

Hypothesis testing, basic structure

Effect exposure **E** on disease **D**

$$\mathbf{E} \longrightarrow \mathbf{D} \quad \text{in population}$$

1. Null hypothesis H_0: no effect (often "$\beta = 0$")
2. Alternative hypothesis H_1: there is an effect (often "$\beta \neq 0$")
3. Calculate value of test statistic *TEST* based on sample data
4. Calculate p-value: probability that *TEST* exceeds some value if H_0 were true
5. If p-value small (often: $< 5\%$): reject H_0
   unlikely that observed difference is due to chance

Hypothesis testing, basic structure

Effect exposure **E** on disease **D**

$$\mathbf{E} \longrightarrow \mathbf{D} \quad \text{in population}$$

1. Null hypothesis H_0: no effect (often "$\beta = 0$")
2. Alternative hypothesis H_1: there is an effect (often "$\beta \neq 0$")
3. Calculate value of test statistic *TEST* based on sample data
4. Calculate p-value: probability that *TEST* exceeds some value if H_0 were true
5. If p-value small (often: $< 5\%$): reject H_0
   unlikely that observed difference is due to chance
6. Otherwise: do not reject H_0
   Does not imply that H_0 is true (power):
   No proof of effect $\neq$ proof of no effect

Outline

Types of study questions

Sampling variation

Binary Variable; Estimating Proportions

Testing a hypothesis

- Single proportion
- Compare proportions between two subgroups
- Categorical variables - more than 2 groups
- Alternative tests for specific settings

2x2 table; example group and outcome

	Disease present	Disease absent	total
male	32	17	49
female	118	127	245
total	150	144	294

2x2 table; example group and outcome

	Disease present	Disease absent	total
male	32	17	49
female	118	127	245
total	150	144	294

- Percentage diseased in the male group:

$$\frac{32}{49} = 65\%$$

- Percentage diseased in the female group:

$$\frac{118}{245} = 48\%$$

2x2 table; example group and outcome

	Disease present	Disease absent	total
male	32	17	49
female	118	127	245
total	150	144	294

- Percentage diseased in the male group:

$$\frac{32}{49} = 65\%$$

- Percentage diseased in the female group:

$$\frac{118}{245} = 48\%$$

2x2 table; example group and outcome

	Disease present	Disease absent	total
male	32	17	49
female	118	127	245
total	150	144	294

- Percentage diseased in the male group:

$$\frac{32}{49} = 65\%$$

- Percentage diseased in the female group:

$$\frac{118}{245} = 48\%$$

- Does disease probability differ by gender?

Hypothesis test

- H_0: probability of disease the **same** in both groups
- H_A: probability of disease **differs** per group
- Can the difference be due to chance, i.e. $H_0 : \pi_F = \pi_M$ holds?
- How do we quantify difference from equal probability?
- "Equal disease probability" is same as saying that group and outcome are unrelated/independent: knowing the group does not help in learning the outcome

Chi-squared test for independence

- For each cell: compare **observed number** (O) with **expected number** (E) under null hypothesis of independence
- Large discrepancy between O and E is an indication that probabilities in both groups differ

Expected count under null hypothesis

	Disease present	Disease absent	total
male	32	17	49
female	118	127	245
total	150	144	294

Expected count under null hypothesis

	Disease present	Disease absent	total
male	32	17	49
female	118	127	245
total	150	144	294

- Proportion male: 49/294 = 0.17
- Proportion diseased: 150/294 = 0.51

Expected count under null hypothesis

	Disease present	Disease absent	total
male	32	17	49
female	118	127	245
total	150	144	294

- Proportion male: 49/294 = 0.17
- Proportion diseased: 150/294 = 0.51

Expected count under null hypothesis

	Disease present	Disease absent	total
male	32	17	49
female	118	127	245
total	150	144	294

- Proportion male: 49/294 = 0.17
- Proportion diseased: 150/294 = 0.51
- Expected number diseased and male under H_0:

$$n \times \hat{P}(\text{male}) \times \hat{P}(\text{Diseased}) = 294 \times (49/294) \times (150/294) = 25$$

Chi-squared statistic calculation (I)

	O	E
Disease present + male	32	25
Disease present + female	118	125
Disease absent + male	17	24
Disease absent + female	127	120
Total	294	294

Chi-squared statistic calculation (II)

	O	E	**O-E**
Dis. present + male	32	25	7
Dis. present + female	118	125	-7
Dis. absent + male	17	24	-7
Dis. absent + female	127	120	7
Total	294	294	0

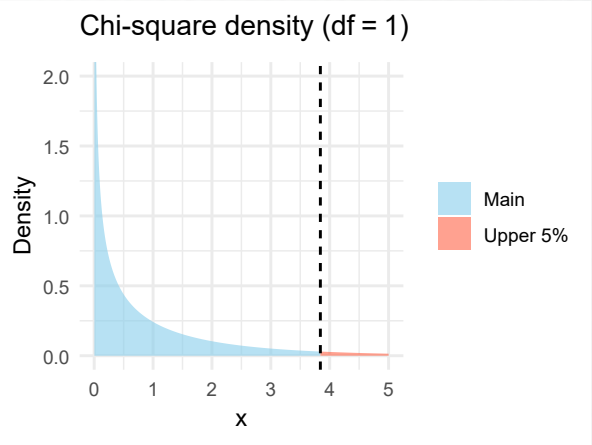
Always equals 0

Chi-squared statistic calculation (III)

	O	E	O-E	$(O-E)^2/E$
Dis. present + male	32	25	7	1.960
Dis. present + female	118	125	-7	0.392
Dis. absent + male	17	24	-7	2.042
Dis. absent + female	127	120	7	0.408
Total	294	294	0	$X^2=4.802$
				P = 0.028

Under null hypothesis, statistic X^2 has a chi-squared distribution with one degree of freedom

Upper percentile of χ^2-distribution



- Starts at zero
- 95-percentile at 3.84

Chi-squared test of independence in R

```
> chisq.test(matrix(c(32,118,17,127),nrow=2),correct=FALSE)        # Variant 1

        Pearson's Chi-squared test
X-squared = 4.802, df = 1, p-value = 0.02843

> prop.test(x=c(32,118),n=c(49,245),correct=FALSE)    # Variant 2, with CI for difference

        2-sample test for equality of proportions without continuity correction
X-squared = 4.802, df = 1, p-value = 0.02843
alternative hypothesis: two.sided
95 percent confidence interval for the difference:
 0.024 to 0.32
sample estimates:
  prop 1   prop 2
    0.65     0.48
```

- Note: correct=TRUE gives “Yates’ continuity correction” (default in R)

Interpreting the chi-squared test of independence

- After identifying the association, you may want to know the strength of this association
- The strength of the association is measured by the difference in proportion, relative risk (RR), or odds ratio (OR) (see Thursday class on logistic regression)

Outline

Types of study questions

Sampling variation

Binary Variable; Estimating Proportions

Testing a hypothesis

- Single proportion
- Compare proportions between two subgroups
- Categorical variables - more than 2 groups
- Alternative tests for specific settings

Chi-squared test for independence

- The chi-squared statistic can be used for testing association in any two-way contingency table
- Variables do not need to correspond to group and outcome
- For a table with c column and r rows, the distribution of the statistics (under null hypothesis) is a chi-squared distribution with $(r-1) \times (c-1)$ degrees of freedom.

2x3 table

Body Mass Index (BMI)

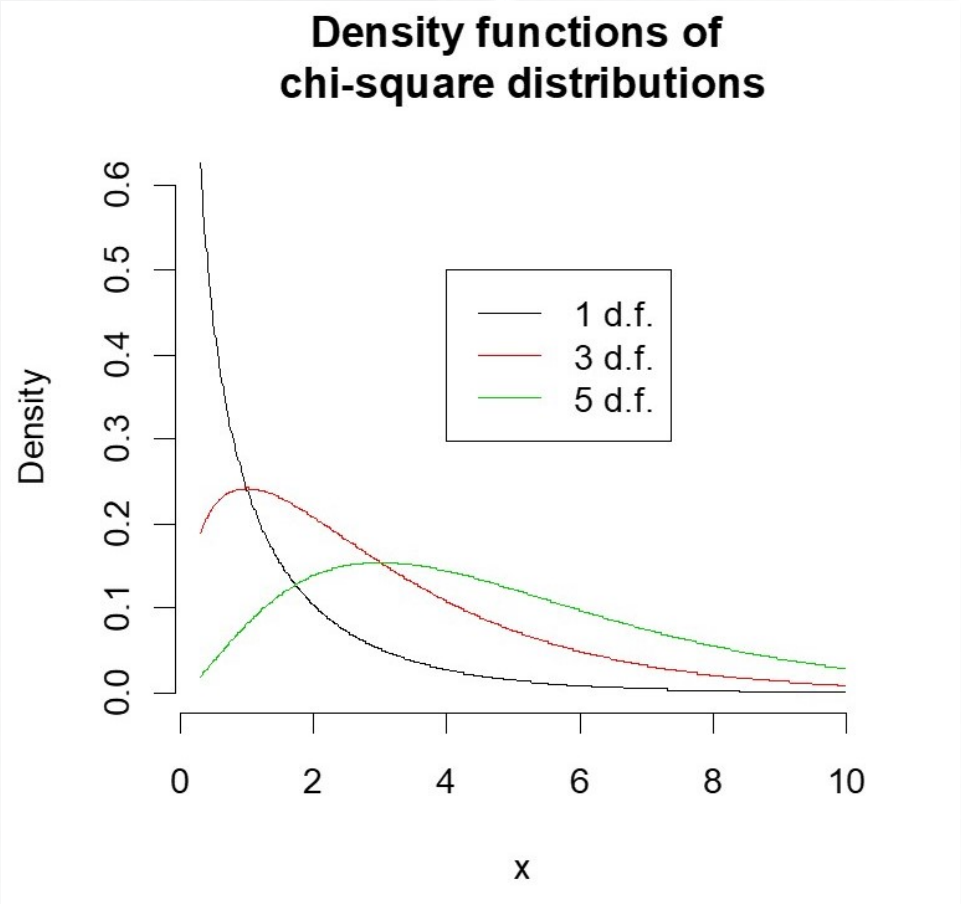
	Low	Normal	Obese	total
male	a	b	c	a+b+c
female	d	e	f	d+e+f
total	a+d	b+e	c+f	

Example with more than two subgroups

TLR2 T597C SNP bacterial genotype frequencies in patients with TBM caused by Beijing genotype isolates compared to chord blood controls (Caws *et al.*; PLoS Pathog 2008)

	Genotype			
Lineage/group	TT	TC	CC	Row Total
Chord blood controls	205 0.544	154 0.408	18 0.048	377
TBM Beijing	24 0.400	25 0.417	11 0.183	60
Column Total	229	179	29	437

Density function of chi-squared distributions



Example with more than two groups in R

```
> chisq.test(matrix(c(205,24,154,25,18,11),ncol=3),
corr=FALSE)

        Pearson's Chi-squared test

data:  matrix(c(205,24,154,25,18,11),ncol=3)
X-squared = 16.3897, df = 2, p-value = 0.0002761
```

"There is clear evidence for an association between TLR2 T597C genotype and TBM caused by Beijing genotype isolates (p=0.0003)."

Outline

Types of study questions

Sampling variation

Binary Variable; Estimating Proportions

Testing a hypothesis

- Single proportion
- Compare proportions between two subgroups
- Categorical variables - more than 2 groups
- Alternative tests for specific settings

Fisher's exact test

- Used in small samples: expected number under null hypothesis is ≤ 1 in at least one cell
  - note: many programs (including R) give a warning if chi-squared test is used and expected frequency under null hypothesis is ≤ 5 in at least one cell

Chi-square trend test (Cochran-Armitage)

- One or both variables are seen as ordinal
  - example: MRC grade at diagnosis (values I, II and III) in HIV-negative and HIV-positive TBM patients

	I	II	III	total
HIV−	298	342	149	702
HIV+	196	251	73	520

Mcnemar test for paired dichotomous outcomes

- Two correlated observations per individual
  - example: disease status per individual before and after treatment

	mild after	severe after	total
mild before	100	50	150
severe before	200	100	300
total	300	150	450