Using Summary Statistics to Learn Probability

YAP Von Bing, Statistics and Applied Probability, NUS
Reference: Statistics 4e by Freedman, Pisani and Purves

May 31, 2012

Topics

- Summary statistics
- The frequency theory of probability theory
- Computer simulation
- Learning outcomes and mathematics syllabi

Summary statistics

- Average, median, proportion, standard deviation, percentile, histogram, plot, chart, ... make complex data comprehensible.
- Statistical modelling makes probabilistic assumptions, some doubtful. For example, a regression model assumes that the response is linearly related to each covariate. Summary statistics should be presented before committing to statistical modelling.

Ideal learning outcomes

- To handle small data sets without a calculator.
- To describe procedure in terms of readily automated operations. Example: the standard deviation of a list of numbers.
  (1) Find the deviations: data minus average.
  (2) Find the root-mean-square of the deviations.
  Simple operations on large data are ideas worth learning.
- To use summary statistics to answer real qualitative and quantitative questions.

What is probability?

- Secondary 2 syllabus: "a measure of chance". Textbooks: "degree of belief". In *Gattaca*, the infant Vincent has a "89% chance of attention deficit disorder". What do they mean?
- The frequency theory: Suppose an experiment can be conducted a large number of times, independently and under identical conditions. The probability (chance) of an event is roughly equal to the proportion of times it is observed.
- According to the theory, with infinitely many experiments, the proportion equals the probability.

Tossing a coin

- A coin can be tossed with the same method independently many times. $P(H) = 0.5$ means: In a large number of tosses, we see heads in about half of them.

- South African mathematician John Kerrich wrote *An experiment introduction to the theory of probability* based on 10,000 coin tosses.

Rolling a die

- A die can be rolled with the same method independently many times. $P(1) = 1/6$: In a large number of tosses, we see one spot in about 1/6 of them.

- What is the chance that the outcome is a multiple of 3? In 6,000 rolls, we get about 1,000 3's and 1,000 6's: about 2,000 multiples of 3. Answer:

$$\frac{2,000}{6,000} = \frac{1}{3}$$

The addition rule

- When rolling a die,

$$P(3 \text{ or } 6) = P(3) + P(6)$$

- More generally, if A and B are mutually exclusive events, then

$$P(A \cup B) = P(A) + P(B)$$

- The addition rule is derived readily in the frequency theory.

Random draws

Make two random draws without replacement from a box containing three identical balls, coloured red, white and blue.

1. What is the chance that the second ball is white?

2. The first ball turns out to be red. Now what is the chance that the second ball is white?

Let $A = \{\text{first ball is red}\}$, $B = \{\text{second ball is white}\}$.
Answers: The unconditional chance $P(B) = 1/3$; the conditional chance $P(B|A) = 1/2$.

Simulation

The computer can be used to simulate the process of drawing without replacement 10,000 times.

1st	2nd
●	●
●	●
●	○
●	○
●	○
⋮	⋮

Table: The first five results of a computer simulation.

	2nd = $\color{red}\bullet$	2nd = $\circ$	2nd = $\color{blue}\bullet$	Row sum
1st = $\color{red}\bullet$	0	1648	1636	3284
1st = $\circ$	1666	0	1660	3326
1st = $\color{blue}\bullet$	1699	1691	0	3296
Column sum	3365	3339	3296	10000

$\text{Prop}(\text{2nd} = \circ) = 3339/10000 \approx 0.334 \approx 1/3.$

Among $(\text{1st} = {\color{red}\bullet})$, $\text{prop}(\text{2nd} = \circ) = 1648/3284 \approx 0.502 \approx 1/2.$

Joint probability

- What is the chance of a red ball followed by a white ball?

  Solution: Imagine repeating the process 6,000 times. In about 2,000 times, the first ball is red. Among these, in about 1,000 the second ball is white. Answer: the joint probability $P(A \cap B) = 1/6$.

- Simulation: Prop(1st $=$ $\bullet$, 2nd $= \circ) = 1648/10000 \approx 0.165$.

- Furthermore, $1/6 = 1/3 \times 1/2$, or $P(A \cap B) = P(A)\, P(B|A)$.

The multiplication rule

- Let A and B be any events (of positive probability). Then

$$P(A \cap B) = P(A)\, P(B|A) = P(B)\, P(A|B) \qquad (1)$$

This can be derived in the frequency theory.

- If $P(B|A) = P(B)$, or equivalently, $P(A|B) = P(A)$, A and B are independent. Then

$$P(A \cap B) = P(A)\, P(B) \qquad (2)$$

- Secondary 3/4, H1, H2: "addition and multiplication of probabilities", "mutually exclusive events and independent events". Seem to refer to (2).

- It is easy for students to forget the independence condition in (2). Safer is the more general (1).

- While (1) is in 9233, H1 and H2 have the equivalent

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

  A picky criticisim: this is like a definition of $P(A|B)$, unnecessary in the frequency theory.

The Kolmogorov axioms

Let Ω be a set (sample space), $\mathcal{F}$ a suitable set of subsets of Ω (events), and $P : \mathcal{F} \to [0,1]$ a function (probability) satisfying

1. $P(\Omega) = 1$.
2. If $E_1, E_2, \ldots \in \mathcal{F}$ are disjoint, then

$$P\left(\bigcup_{n=1}^{\infty} E_n\right) = \sum_{n=1}^{\infty} P(E_n)$$

Given such a $(\Omega, \mathcal{F}, P)$, if $A, B \in \mathcal{F}$, we define the conditional probability

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

Frequency theory and Kolmogorov

- Kolmogorov's axioms need to assume the addition and multiplication rules explicitly. This is a very general abstract theory.
- All the axioms can be derived from the frequency theory. So all consequences of Kolmogorov's axioms also hold in the frequency theory.
- Even though the frequency theory is less general, it is good enough for most statistical applications.

Random variables

- A random variable X is a procedure for generating numbers. X is completely described by its distribution. In the discrete case, this is a list of possible values and the corresponding probabilities.
- The expectation $\mathrm{E}(X)$ and the standard deviation $\mathrm{SD}(X)$ are constants, while X is random. How are they related?

The binomial distribution

- Let X have a binomial distribution with parameters n and p, i.e., X is the number of heads in n independent tosses of a coin with $P(H) = p$. We know that

$$\mathrm{E}(X) = np, \qquad \mathrm{SD}(X) = \sqrt{np(1-p)}$$

- Generate many numbers from the distribution. Their average will be around np, and their SD will be around $\sqrt{np(1-p)}$.

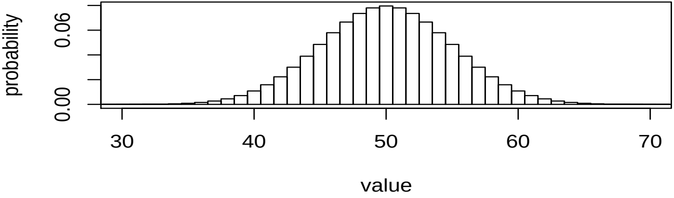
Binomial simulation

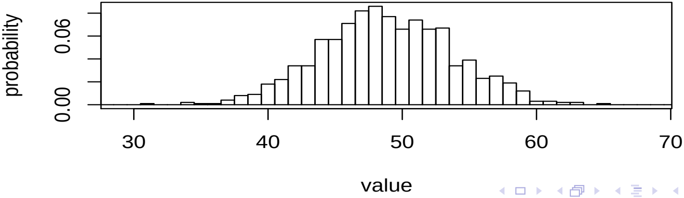
Each distribution is sampled 1,000 times, summarised by the average and SD.

B(n, p)	Average (np)	SD ($\sqrt{np(1-p)}$)
B(25,0.5)	12.51 (12.50)	2.52 (2.50)
B(100,0.5)	50.18 (50.00)	5.05 (5.00)
B(100,0.2)	20.15 (20.00)	3.97 (4.00)
B(100,0.01)	0.95 (1.00)	0.95 (0.10)

The range of B(100,0.5) is 0 to 100, but most of the time the result is within 5 of 50.

Theory vs simulation: B(100,0.5)





Summarising a random variable

- We may say "The $\mathrm{B}(n, p)$ random variable is around np, give or take $\sqrt{np(1-p)}$ or so". If you sample many numbers, their average is around np, the SD is around $\sqrt{np(1-p)}$. Indeed, the histogram is close to the theoretical distribution.

- The statement applies to any random variable, and has the same frequency interpretation. It is analogous to summarising a data set by its average and SD.

Strength of the frequency theory

- Gives precise meaning to probability, both unconditional and conditional, and the expectation, standard deviation and distribution of a random variable.
- Useful for elementary problem solving. *Imagine repeating the process 1,000 times, ...*
- The addition and multiplication rules can be derived.
- Calculations can be checked by simulation. This prepares the ground for the converse: using simulation to estimate difficult probabilities.

Limitation of the frequency theory

- The experiment must be repeatable independently and under the same conditions.
- "Vincent has a 89% chance of attention deficit disorder". The frequency theory is hard to apply, since conditions change by day, and days may not be independent. Maybe Vincent is part of a large group of similar people, though it is not clear how this population can be identified.
- The frequency theory does not apply in every situation. Knowing the limit of a tool is important.

Conclusion

- Learning outcomes for summary statistics should be geared towards effective handling of large data sets.
- Summary statistics help clarify probability concepts and rules. This is possible in the frequency theory, but may not be in others.
- The general multiplication rule is preferable to the special case for independent events.
- Tables and diagrams presented here are made with the R language. The scripts are available upon request. Online probability applets can also be useful.

Logically equivalent questions

In a Mathematical Statistics class in 2011:

- Pre-test: The unconditional probability of event A is $1/3$; the unconditional probability of B is $1/2$. If A and B are independent, they must also be mutually exclusive. **False.**
- Midterm test: The unconditional probability of event A is $1/2$. The unconditional probability of event B is $1/3$. If A and B are mutually exclusive, they cannot be independent. **True.**
- Final examination: Same as pre-test.

After first two tests, answers were given, but content not discussed. Probability is a pre-requisite.

Results

	Pre	Midterm	Final
% correct	90	63	86

- Reasoning is more than knowing whether a statement is true.
- Knowledge fades without explicit reinforcement.

Who learnt from the midterm?

	0	1	Pre-test
0	5 (9%)	11 (20%)	16 (29%)
1	2 (4%)	38 (68%)	40 (71%)
Final	7 (12%)	49 (88%)	56 (100%)

Table: Students who got midterm wrong. Rows: pre-test; columns: final.

	0	1	Pre-test
0	2 (2%)	3 (3%)	5 (5%)
1	6 (6%)	86 (89%)	92 (95%)
Final	8 (8%)	89 (92%)	97 (100%)

Table: Students who got midterm right. Rows: pre-test; columns: final.