Probability & Statistics - Important Ideas

(10 pages; 28/6/21)

See also "Useful Results - Statistics"

Contents

(A) Variance

(1) Variance of a random variable

$Var(X) = E[(X-\mu)^2]$

$= E[X^2 - 2X\mu + \mu^2]$

$= \sum_x (x^2 - 2x\mu + \mu^2)P(X = x)$

$= [\sum_x x^2 P(X = x)] - [2\mu \sum_x xP(X = x)] + \mu^2 \sum_x P(X = x)$

$= E(X^2) - 2\mu E(X) + \ \mu^2$

$= E(X^2) - 2\mu^2 + \mu^2$

$= E(X^2) - \mu^2$

(2) Sample variance

The recommended version to use is the one with the denominator of $n-1$ (the 'corrected' version). This is an unbiased estimate for the population variance, but is recommended to be used, whether or not the population variance is mentioned. Arguably, the uncorrected version (with a denominator of n) is the variance of the sample, but for the 'sample variance' it's best to use the corrected version. See (3) for the proof of its being an unbiased estimate.

$s^2 = \frac{1}{n-1}\sum(x_i - \overline{x})^2$

$= \frac{1}{n-1}\left\{\sum\left(x_i{}^2 - 2x_i\overline{x} + \overline{x}^2\right)\right\}$

$= \frac{1}{n-1}\left\{(\sum x_i{}^2) - 2\overline{x}(\sum x_i) + n\overline{x}^2\right\}$

$= \frac{1}{n-1}\left\{(\sum x_i{}^2) - 2\overline{x}(n\overline{x}) + n\overline{x}^2\right\}$

$= \frac{1}{n-1}\{(\sum {x_i}^2) - n\overline{x}^2\}$

(3) The random variable $S^2 = \frac{1}{n-1}([\sum X^2] - n\overline{X}^2)$ is an unbiased estimator for the population variance σ^2.

Proof

$E(S^2) = \frac{1}{n-1}\{[\sum E(X^2)] - nE(\overline{X}^2)\}$

$= \frac{1}{n-1}\{nE(X^2) - nE\left(\overline{X}^2\right)\}$

$= \frac{n}{n-1}\{E(X^2) - E\left(\overline{X}^2\right)\}$

Now $\sigma^2 = E(X^2) - \mu^2$, so that $E(X^2) = \sigma^2 + \mu^2$

Also, $Var(\overline{X}) = E\left(\overline{X}^2\right) - \mu^2$,

and $Var(\overline{X}) = Var(\frac{X_1+\cdots+X_n)}{n} = \frac{1}{n^2}Var(X_1 + \cdots + X_n)$

$= \frac{1}{n^2}\left(nVar(X_i)\right) = \frac{\sigma^2}{n}$,

so that $E\left(\overline{X}^2\right) = \frac{\sigma^2}{n} + \mu^2$

Then $E(S^2) = \frac{n}{n-1}([\sigma^2 + \mu^2] - \left[\frac{\sigma^2}{n} + \mu^2\right])$

$= \frac{n\sigma^2}{n-1}(1 - \frac{1}{n})$

$= \frac{n\sigma^2}{n-1}(\frac{n-1}{n})$

$= \sigma^2$

(4) $Var(aX + b) = a^2VarX$

Proof

$Var(aX+b)=E[(aX+b)^2]-[E(aX+b)]^2$

$=E[a^2X^2+2abX+b^2]-[aE(X)+b]^2$

$=a^2E(X^2)+2abE(X)+b^2-[a^2[E(X)]^2+2abE(X)+b^2]$

$=a^2E(X^2)-a^2[E(X)]^2$

$=a^2VarX$

(5) If $Var(X_1)=Var(X_2)$, and $Y=X_1+X_2$,

then $Var(Y)=Var(X_1)+Var(X_2)=2Var(X_1)$,

provided that X_1& X_2 are independent

But if $Z=2X_1$, then $Var(Z)=4Var(X_1)$.

(6) Where X & Y are not necessarily independent:

$Var(aX\pm bY)=a^2VarX+b^2VarY\pm 2abCov(X,Y)$,

where $Cov(X,Y)=E(XY)-E(X)E(Y)$

(7) Useful device for determining $Var(X)$:

$E(X^2)=E[X(X-1)]+E(X)$ [as $r(r-1)$ divides into $r!$]

(B) Data

(1) Quartiles

To determine Q_1: take the items to the left of the median (or, if the median is the average of x_r & x_{r+1}, take the items up to and including x_r), and obtain their median. Similarly for Q_3.

(C) Independence and Conditional Probability

(1) By considering regions in a Venn diagram, $P(B|A) = \frac{P(A\cap B)}{P(A)}$.

Alternatively, $P(A\cap B) = P(A)P(B|A)$.

(2) A and B may or may not occur simultaneously. Examples are:

(a) A = pupil is late to school on Monday; B = pupil is late to school on Tuesday

(b) A = it is raining in London; B = it is raining in New York

The formula $P(A\cap B) = P(A)P(B|A)$ is easier to understand if A occurs before B, but is also valid if A and B take place at the same time. In the case of (b), for example, we could first of all consider the weather in London, and then the weather in New York (given that it is raining in London).

(3) Events A and B are independent when A doesn't influence B (and vice-versa), so that $P(B|A) = P(B)$ (1)

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

so that $P(A\cap B) = P(A)P(B)$ (2)

Either (1) or (2) can be used as a test for independence.

(4) Sometimes independence isn't immediately obvious. For example, suppose that the following 2-way table applies in connection with pupils choice of subjects:

	Italian	German	Spanish	Total
Male	18	14	8	40
Female	23	21	16	60
Total	41	35	24	100

Then $P(M) = \frac{40}{100} = \frac{2}{5}$ and $P(M|S) = \frac{8}{24} = \frac{1}{3}$

so that $P(M|S) \neq P(M)$ and hence M and S aren't independent.

But $P(M|G) = \frac{14}{35} = \frac{2}{5}$, so that M and G are independent.

Alternatively, $P(S) = \frac{24}{100} = \frac{6}{25}$, $P(M \cap S) = \frac{8}{100} = \frac{2}{25}$, and

$P(M).P(S) = \frac{2}{5} . \frac{6}{25} = \frac{12}{125}$, and so M and S aren't independent, as

$P(M \cap S) \neq P(M).P(S)$

Also, $P(G) = \frac{35}{100} = \frac{7}{20}$, $P(M \cap G) = \frac{14}{100} = \frac{7}{50}$, and

$P(M).P(G) = \frac{2}{5} . \frac{7}{20} = \frac{7}{50}$, and so M and G are independent, as

$P(M \cap G) = P(M).P(G)$

(5) Mutually exclusive events

This shouldn't be confused with independence.

In the above example, G and S are mutually exclusive, as
$P(G \cap S) = 0$

But G and S are not independent, as $P(G|S) = 0 \neq P(G)$.

Alternatively, $P(G \cap S) = 0 \neq P(G).P(S)$

(D) Hypothesis Tests

When an extreme value occurs, there are two possible explanations: (a) H_0 is correct, and the extreme value occurred by chance, and (b) H_0 is not correct.

The hypothesis test establishes whether explanation (a) is too implausible to be the correct one.

If H_0 really is true, there will still be occasions when implausibly extreme values occur, but we are prepared to be wrong 5% of the time (if the significance level is 5%). For a one-tailed test, this will happen when H_0 is true and the value is greater than or equal to some critical value (CV), such that $P(X \geq CV) = 0.05$

(E) Poisson distribution

(1) Each Poisson distribution is associated with an interval over which events occur (usually a time interval, but sometimes distance). Thus the Poisson parameter λ will change if the interval is changed.

(2) Conditions for a Poisson model to be appropriate:

(i) random & independent events

[Note: These are invariably treated as a single idea for exam purposes. It is however possible for events to be random but not independent, and vice-versa.

Example A: In a football match, where the variable is the number of goals scored in a match (by either side), one team may try harder to score a goal if the other side has just scored; ie events wouldn't be independent, but would still be random.

Example B: If the variable is the number of buses passing a particular point in an hour, then the existence of a timetable would make the events non-random, but they could still be independent (with one event not influencing another).

(ii) constant rate

This means that the rate of an event occurring doesn't change over the interval in question.

Note: There is a common misconception here. Data is often used to establish the Poisson parameter λ, and this data might be (unwisely) collected over too long a period, with the rate of occurrence changing over the period. (For example, the rate of outbreak of fires is likely to vary considerably over a 24 hour period.) The requirement for a constant rate does not refer to behaviour over the data collection period (though a varying rate over this period would call into question the appropriateness of the value of λ being used).

(iii) If a sample of outcomes is available, then the sample mean should be close to the sample variance, if a Poisson model is appropriate.

(F) Continuity Correction

(when a discrete variable is being approximated by a Normal variable)

There is a correspondence between the discrete values and the continuous values of the Normal distribution,

so that if X is discrete and Y is the Normal approximation, then $10 \leq X \leq 20 \leftrightarrow 9.5 \leq Y < 20.5$,

as the Y interval must include all values that round to 10 or 20; though we can also write $9.5 < Y < 20.5$, as $P(Y = 9.5) = 0$ for a continuous variable

Examples

(i) $X \geq 20 \ \leftrightarrow \ Y > 19.5$

(ii) $X < 10 \equiv X \leq 9 \ \leftrightarrow \ Y < 9.5$

(G) Counting

(1) Selections

(i) Ordered selections with repetition

Number of ways of selecting r items from n, if repetitions are allowed, and order is important $= n^r$

(ii) Ordered selections without repetition

Number of ways of selecting r items from n, if repetitions are not allowed, and order is important

$= n(n-1)\dots(n-[r-1]) = n(n-1)\dots(n-r+1)$

[Known as a Permutation]

$P(n,r)$ or ${}^nP_r = \frac{n!}{(n-r)!} = n(n-1)\dots(n-r+1)$

(iii) Unordered selections without repetition

Number of ways of selecting r items from n, if repetitions are not allowed, and order is not important

[Known as a Combination.]

$C(n,r)$ or ${}^nC_r$ or $\binom{n}{r} = \frac{n!}{r!(n-r)!} = \frac{n(n-1)\ldots(n-r+1)}{r!}$

[${}^nC_r$ can be obtained from ${}^nP_r\ = \frac{n!}{(n-r)!}$ by dividing by $r!$, to remove duplication (the ${}^nP_r$ ordered ways can be divided into groups of $r!$, containing the same items, but in a different order).]

(iv) Unordered selections with repetition [MAT/STEP]

Number of ways of selecting r items from n, if repetitions are allowed, and order is not important

eg $BBCE$ selected from $ABCDEF$ $(r = 4, n = 6)$

write as $|XX|X||X|$

(| indicates that we are moving on to the next letter, and XX indicates that we are selecting 2 items from the current letter: so $|XX|X||X|$ means: move on to B (without selecting any As); then select 2 Bs; then move on to the Cs; select 1 C; move on to D, and then on to E; select 1 E; then move on to F, but select no Fs)

= Number of ways of choosing r positions for the Xs,

out of the $n-1$ $|s$ and r Xs (giving a total of $n-1+r$)

$= \binom{n-1+r}{r}$