S3

Will Bevington Callum O'Brien Alex Pace

December 15, 2015

Contents

1 Combining Random Variables

Let X and Y be two independent random variables with means $\mathrm{E}(X)$ and $\mathrm{E}(Y)$ respectively, and variances $\mathrm{Var}(X)$ and $\mathrm{Var}(Y)$ respectively;

$$\mathrm{E}(X \pm Y) = (\mathrm{E}X) \pm \mathrm{E}(Y)$$

$$\mathrm{Var}(X \pm Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$$

$$\mathrm{E}(aX \pm b) = a\mathrm{E}(X) \pm b$$

$$\mathrm{Var}(aX \pm b) = a^2\mathrm{Var}(X) + b$$

The latter two formulae should be recalled from S1. We can combine these to acquire:

$$\mathrm{E}(aX \pm bY) = a\mathrm{E}(X) \pm b\mathrm{E}(y)$$

$$\mathrm{Var}(aX \pm bY) = a^2\mathrm{Var}(X) + b^2\mathrm{Var}(Y)$$

Additionally, the combination of two independent normal distributions is also a normal distribution;

$$X \ N(\mu_x, \sigma_x^2)$$

$$Y \ N(\mu_y, \sigma_y^2)$$

$$(aX \pm bY) \ N(a\mu_x \pm b\mu_y, a^2\sigma_x^2 + b^2\sigma_y^2)$$

2 Sampling Frames

Population The whole set of items that are of interest.

Census Observes or measures every memeber of a population.

Sample Survey A selection of observations taken from a subset of the population which is used to find out information about the population as a whole.

Random Sample A sample in which every possible sample of size n has an equal chance of being selected.

Sampling Frame A list identifying every single sampling unit that could be included in the sample

2.1 Random Sampling

Random Number Sample

Give each sampling unit in the sampling frame a number and use a random number generator or random number tables to select required number of sampling units.

Lottery Sample

Put the sampling units from the sampling frame into a "hat" and select randomly without replacing.

Positives

- Random and free from bias
- Easy to carry out

Negatives

- Not suitable for large sample sizes

2.2 Systematic Sampling

Pick at required intervals from an ordered list, e.g. I want a sample of 15 from 60: $\frac{60}{15} = 4$ therefore choose a starting point randomly from one of the first four sampling unit from the ordered list, then choose every fourth sampling unit after until you have selected 15.

Positives

- Suitable for large samples
- Is easy to carry out

Negatives

- Sample is not random unless the ordered list is random
- Can introduce bias

2.3 Stratified Sampling

A form of random sampling: The population is split into mutually exclusive groups (strata). Random samples are taken from each strata, the relative size of each corresponds to the same ratio as each strata's representation in the total population.

Positives

- Works well with large samples that can be split into mutually exclusive groups
- Reflects a populations structure

Negatives

- Takes longer than random sampling
- Within each strata the problems are the same as with any random sample.
- Ill defined strata can overlap (meaning they are no longer mutually exclusive)
- Can't provide accurate data when strata overlap

2.4 Quota Sampling

When no sampling frame is available, quota sampling may be used. The population is divided into groups (as with stratified sampling). Quotas for each group are created that corrrespond with the groups representation in the total population. The interviewer then selects sampling units until each quota is reached.

Positives

- Administering the test is easy
- Test is low cost
- Test is quick if the sample is small

Negatives

- Introduces interviewer bias
- Can't estimate sampling errors

3 Types of Data

3.1 Primary Data

When you collect data, or someone collects data on your behalf.

Positives

- You have control over the type and method of collection
- The exact data needed is collected
- The Accuracy is known

Negatives

- Expensive (money and time)

3.2 Seconday Data

Second hand data, collected by another person or organisation.

Positives

- Cheaper than gathering primary data (time and money)
- Large amounts of data are easily available on the internet
- Access to data over time (trends)

Negatives

- Bias is not always acknowledged
- Accuracy is not known
- Certain data can be in a form that is difficult to deal with

4 Estimating Population Parameters using a Sample

A statistic which is used to estimate a population parameter is called an estimator. A particular value is called an estimation. If X is a random variable then $\mathrm{E}(X)$ would be an estimator of the mean. If a statistic T is an estimator for a population parameter θ and $\mathrm{E}(T) = \theta$ then T is an unbiased estimator for θ. Otherwise, the bias of T is given by the expression

$$\mathrm{E}(T) - \theta$$

Estimators for population parameters can be written using "hat notation," wherein an estimator for a population parameter θ is denoted by $\hat{\theta}$.

$$\bar{X} = \frac{1}{n}\sum_i X_i \Rightarrow \mathrm{E}(\bar{X}) = \mu_X \tag{1}$$

$$S^2 = \frac{1}{n-1}\left(\sum_i X_i^2 - n\bar{X}^2\right) \Rightarrow \mathrm{E}(S^2) = \sigma_X^2 \tag{2}$$

Proof of (1) assuming $E(X+Y) = E(X) + E(Y)$,

$$E(\bar{X}) = E\left(\frac{1}{n}\sum_i X_i\right) = \frac{1}{n}E\left(\sum_i X_i\right)$$

$$E(\bar{X}) = \frac{1}{n}\sum_i E(X_i) = \frac{1}{n}n\mu = \mu$$

4.1 Standard Error

If $\bar{x}$ is an estimator of the mean, $\frac{\sigma}{\sqrt{n}}$ is the standard error. As we probably don't know σ, use S instead $\left(\frac{S}{\sqrt{n}}\right)$. Note that as n increases, the standard error decreases.

4.2 The Central Limit Theorem

The central limit theorem states that if $X_1, X_2, \cdots, X_n$ is a random sample of size n from population with mean μ and variance σ^2 and n is large, then $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$.

Worked Example

Question: A die $\{1, 1, 1, 3, 3, 6\}$ is rolled 40 times and the mean of 40 rolls is calculated. Find an approximation for the probability that $mean > 3$.

Answer:

$$\begin{array}{cccc} x & 1 & 3 & 6 \\ P(X=x) & \frac{1}{2} & \frac{1}{3} & \frac{1}{6} \end{array}$$

$$E(X) = \sum xP(X=x) = \frac{5}{2}$$

$$E(X^2) = \sum x^2P(X=x) = \frac{19}{2}$$

$$\therefore Var(X) = \frac{19}{2} - \left(\frac{5}{2}\right)^2 = \frac{38-25}{4} = \frac{13}{4}$$

Thus, by the central limit theorem,

$$\bar{X} \sim N\left(\frac{5}{2}, \frac{13}{4} \times \frac{1}{40}\right)$$

$$\sim N\left(2.5, 0.08125\right)$$

$$P\left(\bar{X} > 3\right) = P\left(Z > \frac{3-2.5}{\sqrt{0.08125}}\right) = P\left(Z > 1.7184\right) = 0.0397$$

4.3 Calculating Confidence Intervals for a Population Interval

Typical confidence levels are 99% or 95%. A 95% confidence interval is the range of outcomes that 95% of your results will fall in. A 95% confidence interval for μ from a sample of size n from a population with mean μ and variance σ^2 is:

$$\bar{X} \pm 1.96 \times \frac{\sigma}{\sqrt{n}}$$

A 99% confidence interval of μ is:

$$\bar{X} \pm 2.58 \times \frac{\sigma}{\sqrt{n}}$$

5 Testing Hypotheses

5.1 Differences of Means of two Independent Normal Distributions

$$X \sim N\left(\mu_X, (\sigma_X)^2\right)$$

$$Y \sim N\left(\mu_Y, (\sigma_Y)^2\right)$$

$$X - Y \sim N\left(\mu_X - \mu_Y, (\sigma_X)^2 + (\sigma_Y)^2\right)$$

Taking a sample of n_X from X and n_Y from Y to get $\bar{X}, \bar{Y}$,

$$\bar{X} - \bar{Y} \sim N\left(\mu_X - \mu_Y, \frac{(\sigma_X)^2}{n_X} + \frac{(\sigma_Y)^2}{n_Y}\right) \tag{3}$$

This gives the test statistic,

$$Z = \frac{\left(\bar{X} - \bar{Y}\right) - (\mu_X - \mu_Y)}{\sqrt{\frac{(\sigma_X)^2}{n_X} + \frac{(\sigma_Y)^2}{n_Y}}}$$

If X and Y weren't normally distributed, (3) would still be a good approximation if n_X and n_Y were large (by the central limit theorem.)

Example

$$H_0 : \mu_X = \mu_Y,\ H_1 : \mu_X > \mu_Y,\ \alpha = 0.05$$

	X	Y
sample mean	48	45
σ	5	8
n	25	30

$$Z = \frac{\bar{X} - \bar{Y} - (0)}{\sqrt{\frac{25}{25} + \frac{64}{30}}} = \frac{3}{\sqrt{3.1333}} = 1.6947$$

$$P\left(Z < a\right) = 0.95 \Rightarrow a = 1.6449,\ 1.6947 > 1.6449$$

$\therefore$ There is sufficient evidence to reject H_0 in favour of H_1

6 Goodness of Fit & the χ^2 Distribution

$$
\begin{aligned}
X^2 &= \sum_i \frac{(O_i - E_i)^2}{E_i} \\
&= \sum_i \frac{(O_i)^2 - 2O_iE_i + (E_i)^2}{E_i} \\
&= \sum_i \left(\frac{(O_i)^2}{E_i} - 2O_i + E_i \right) \\
&= \sum_i \frac{(O_i)^2}{E_i} - 2\sum_i O_i + \sum_i E_i \\
\sum_i O_i &= \sum_i E_i = N
\end{aligned}
$$

hence

$$X^2 = \sum_i \frac{(O_i)^2}{E_i} - N$$

where

$$N = N.o.\ cells$$

X^2 is approximated well by χ^2 if none of the expected values fall below five.

6.1 Degrees of Freedom

6.2 Binomial Distribution

If any $E_i < 5$, these cells, along with the corresponding O_i must be merged to avoid this. Then,

$$N.o.\ degrees\ of\ freedom = N - 2$$

if p is estimated by calculation and

$$N.o.\ degrees\ of\ freedom = N - 1$$

otherwise.

6.3 Poisson Distribution

6.4 Uniform Continuous Distribution

6.5 Normal Distribution

We may suspect that some data are normally distributed if they exhibit the characteristic 'bell-shaped' curve and/or if roughly $^2/_3$ of the data lie within one standard deviation from the mean.

A sample from a population $X \sim N\left(\mu, \sigma^2\right)$ consisting of n fields has ν degrees of freedom where

$$\nu = \begin{cases} n-3 & \text{if } \mu \text{ and } \sigma^2 \text{ are estimated} \\ n-2 & \text{if } \mu \text{ or } \sigma^2 \text{ is estimated} \\ n-1 & \text{otherwise} \end{cases}$$

Question (tell me what you think about me)

During an observation on the height of 200 male students the following data were observed:

Height / cm	Frequency
150-154	4
155-159	6
160-164	12
165-169	30
170-174	64
175-179	52
180-184	18
185-189	10
190-194	4

1. Test at the 0.05 level to see if the height of male students could be modelled by a normal distribution with mean 172 and standard deviation 6.
2. Describe how you would modify this test if the mean and variance were unknown.

6.6 Contingency Tables

A contingency table is a way to test whether two variables are independent or linked. The null hypothesis is always that the variables are independent.

Example: Link between School and Grade

Grade	A	B	C	Total
School X	18	12	20	50
School Y	26	12	32	70
Total	44	24	52	120

Table 1: Observed values

$$H_0: \text{ School and grade are independent.}$$

$$H_1: \text{ School and grade are linked.}$$

Assuming H_0, we can calculate expected values by multiplying relevant probabilities and quantities, as

$$A \text{ and } B \text{ are independent} \Leftrightarrow P(A \wedge B) = P(A) \times P(B)$$

giving the following expected values:

hence, for these data,

$$X^2 = 0.92$$

Grade	A	B	C
School X	18.33	10.00	21.67
School Y	25.67	14.00	30.33

Table 2: Expected Values

As these data have two degrees of freedom, the critical value (at significance 0.005) is given by

$$\chi_2^2\left(0.05\right) = 5.99$$

which is greater than X^2. There is therefore insufficient evidence to reject H_0, i.e. school and grade are independent.