Point and Interval Estimation for Categorical Data

MAST90044

Thinking and Reasoning with Data

Dr Julia Polak

Chapter 3 – Lectures 6 and 7

Department of Mathematics & Statistics
The University of Melbourne

Outline

Statistical inference

Statistical inference — drawing conclusions about a *population* from *data*.

Inference involves two main activities:

- Estimation;
- Hypothesis testing.

Estimation involves:

- Point estimation;
- Interval estimation.

Random variables

A *random variable* is a numerical outcome of a random phenomenon.

Random variables can be either:

- Discrete, e.g. count;
- Continuous, e.g. measurement.

Notation: $X \stackrel{\mathrm{d}}{=} \mathbb{D}, \quad X \stackrel{\mathrm{d}}{\approx} \mathbb{D}, \quad \ldots$

Correction for continuity

is an adjustment that is made when a discrete distribution is approximated by a continuous distribution.

Intervals

(upper-case denotes random, lower-case denotes constant = non-random)

$\Pr(a < X < b) = p$ probability interval

(a fixed interval which will contain the value of the RV X with probability p)

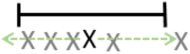
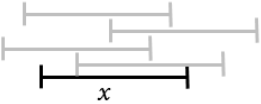
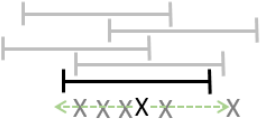
$\Pr(A < x < B) = p$ confidence interval

(a random interval which will contain the parameter x with probability p. Realisation of the random interval gives a confidence interval.)

$\Pr(A < X < B) = p$ prediction interval

(a random interval which will contain a future observation of the random variable X with probability p)

Intervals

		interval	*value*
Probability interval $\Pr(a < X < b) = p$		fixed	RV
Confidence interval $\Pr(A < x < B) = p$		RV	fixed parameter
Prediction interval $\Pr(A < X < B) = p$		RV	RV

Random variables - Binomial

In the long run of repeated samples, the value of a random variable X can be modelled by a rule of probability, which defines the *probability distribution* of the random variable.

An important discrete random variable:

For repeated independent processes with a binary outcome, the number of "successes" X can be described by the *binomial distribution*:

$$X \stackrel{\mathrm{d}}{=} \mathrm{Bi}(n, p)$$

The random variable X represents the number of successes of n independent trials, each with probability of success p.

Example: Roll a die 6 times, count the number of sixes.

$$X \stackrel{\mathrm{d}}{=} \mathrm{Bi}(6, \tfrac{1}{6})$$

Representations of probability distributions

pmf/pdf the probability mass function $p_X(x) = \Pr(X = x)$
[probability density function, $f_X(x)dx = \Pr(X \approx x)$.]
In R, for the $***$ pmf, write `d***()`.
e.g. `dbinom(3,10,0.25)`

cdf the cumulative distribution function $F_X(x) = \Pr(X \leqslant x)$
In R, for the $***$ cdf, write `p***()`.
e.g. `pbinom(3,10,0.25)`

- The cdf function is invertible.
  In R, for the $***$ inverse-cdf, write `q***()`.
  e.g. `qbinom(0.6,10,0.25)`
- $\mathrm{E}(X)$ denotes the *expectation* (mean) of X.
- $\mathrm{var}(X)$ denotes the variance of X.

Moments of probability distributions

Moment

A *moment* is a quantitative measure that gives us the information about the location and shape of the distribution of the random variable.

The probability distribution of X is characterized by its first two moments; the *expected value* (mean) and a *variance* (standard deviation squared)

Expected value

is the mean (μ) of random variable X.

> *a point estimate of how we expect X to behave on-average over the long run*

Note: the symbols $E(X) = \mu$ are used interchangeably.

Variance

is the expected (or average) squared distance (or deviation) from the mean; the *standard deviation squared*, σ^2

Moments of probability distributions

$\mathrm{E}(X)$ and $\mathrm{var}(X)$ are the first two *moments* of the distribution of X.

If $X \stackrel{\mathrm{d}}{=} \mathrm{Bi}(n, p)$,

- $\mathrm{E}(X) = np$
- $\mathrm{var}(X) = np(1-p)$

$\dfrac{X}{n}$ is an estimator of p: $\qquad \hat{p} = \dfrac{X}{n}$

$$\mathrm{E}(\hat{p}) = \mathrm{E}(\frac{X}{n}) = p \quad \text{and} \quad \mathrm{var}(\hat{p}) = \mathrm{var}(\frac{X}{n}) = \frac{p(1-p)}{n}.$$

Point estimation

What is the best estimate of a population parameter, given

- an assumed probability distribution, and
- a sample of data?

- Method of Moments (MM)

  Estimate the population moments using the sample moments.

- Maximum Likelihood (ML)

  Estimate the population parameters by the values that make the sample as probable as possible.

In simpler situations, they give the same answer.

Method of moments

Example:

$$X \stackrel{\mathrm{d}}{=} \mathrm{Bi}(n, p)$$

MM: estimated proportion = observed proportion

$\Rightarrow$ MM estimate is $\hat{p} = x/n$

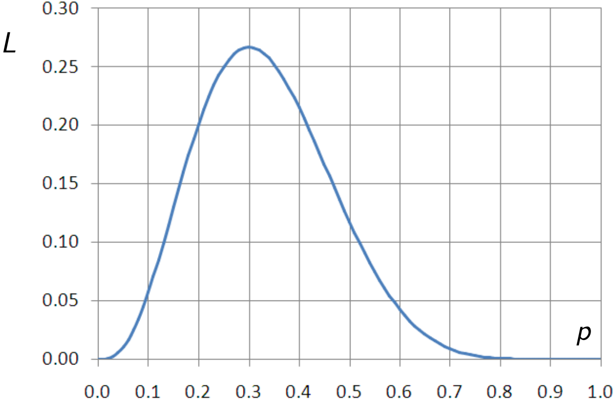
Example: 10 tosses of a coin, 3 heads. $\hat{p} = \frac{3}{10}$

Maximum likelihood

Estimate the population parameters by the values that make the sample as probable as possible.

$$\text{likelihood} = L(p \mid \mathrm{data}) = \Pr(\mathrm{data} \mid p) = \Pr(X = 3 \mid p)$$

$$X \stackrel{\mathrm{d}}{=} \mathrm{Bi}(n, p);\ n = 10, x = 3; \quad \Rightarrow \quad L(p) = \binom{10}{3} p^3 (1-p)^7$$



Interval estimation – Definition

An interval estimate is a set of plausible values for the parameter. More precisely, an interval estimator is a random variable which is expected (with specified probability) to contain the unknown parameter. An *interval estimate* is a *realisation* of an *interval estimator*.

Interval estimation

Confidence intervals - Interpretation

Non-technical definition:
An interval within which we are (95%) confident that the true value of the parameter lies.
Example: a 95% confidence interval for a population proportion p is an interval within which we are 95% sure that the true value of p lies.

Technical definition:
The confidence coefficient (e.g. 95%) is the long-term percentage of such intervals containing the true value.

A confidence interval is a realisation of a *random interval*: it is different for every sample from the population, and may or may not contain the true value.

A confidence interval gives a measure of the precision of the parameter estimate.

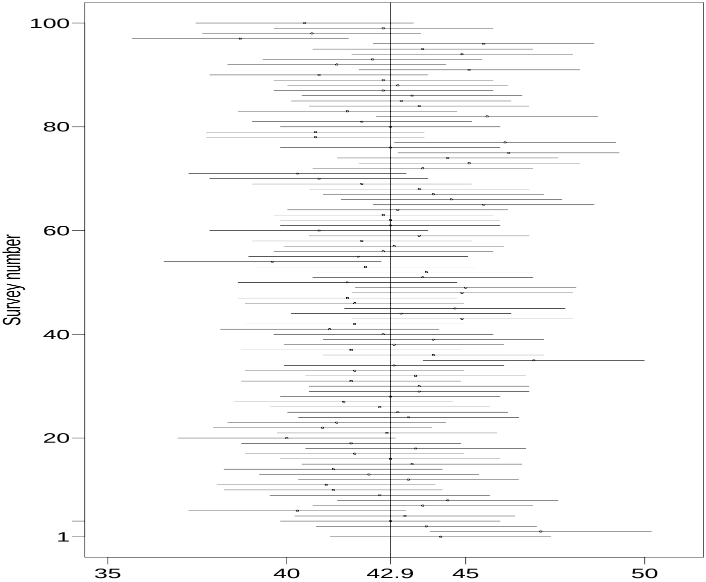
Interval estimation

When sample size increases (n↑)

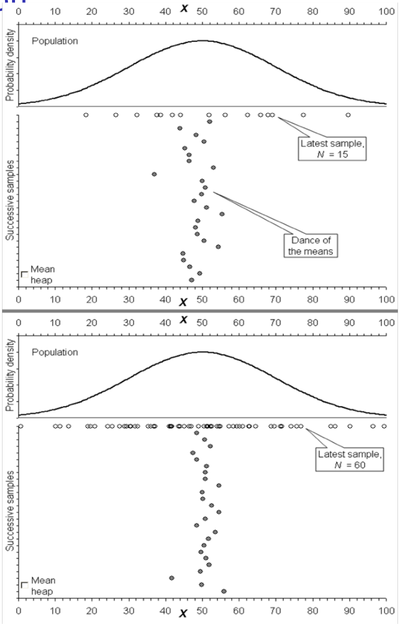
Confidence interval become narrower

Mean is “dance” less

Interval estimation

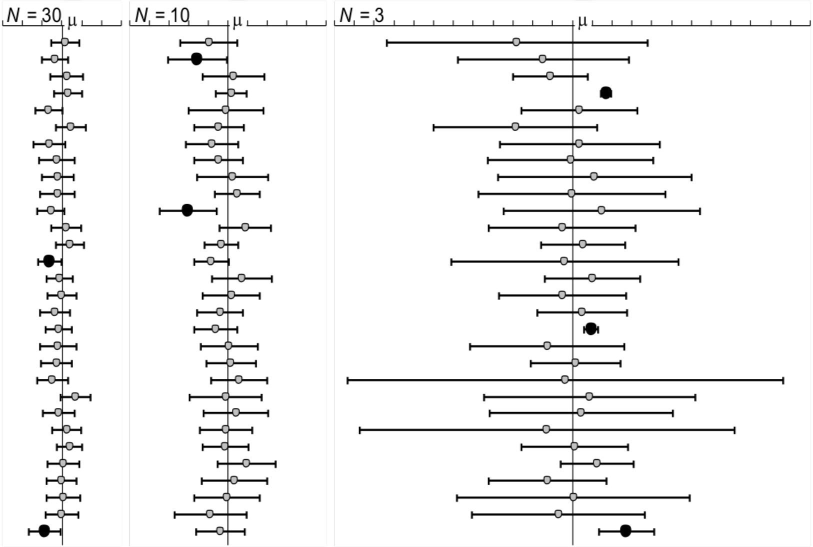


Interval estimation



Interval estimation - 30 samples of size N, 95% CIs.

1.5 ($\approx$ 5%) of 30 samples will *not* contain μ.



Interval estimation of proportions

There are many methods: here are three that are well used:

1. Wald (first approximation);
2. Wilson (second approximation, correction for continuity);
3. Clopper-Pearson ("exact").

Wald confidence intervals (first approximation)

If n is sufficiently large, the sampling distribution of the sample proportion is approximately normal:

$$\hat{P} \stackrel{\mathrm{d}}{\approx} \mathrm{N}\left(p,\ \sqrt{\frac{p(1-p)}{n}}\right)$$

($\hat{P} = X/n$ is the random variable, $\hat{p}$ is the realisation.)

A 95% *Wald* confidence interval for p is

$$\hat{p} \pm 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

i.e. $\quad \mathrm{est} \pm$ "2" se

Wald confidence intervals (first approximation)

A 95% *Wald* confidence interval is

$$\hat{p} \pm 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Example: proportion of National Party voters in an electorate

Random sample of 300 people: 142 NP voters.

$\hat{p} = 142/300 = 0.473.$

$$0.473 \pm 1.96\sqrt{\frac{0.473(1-0.473)}{300}} = 0.473 \pm 0.056 = (0.417, 0.529)$$

Example: proportion of left-handed women

Random sample of 30 women: 2 left-handed. $\hat{p} = 2/30 = 0.067.$

$$0.067 \pm 1.96\sqrt{\frac{0.067(1-0.067)}{30}} = 0.067 \pm 0.089 = (-0.022, 0.156)$$

Improvement needed!

Wilson confidence intervals (second approximation)

- Wilson score interval insure that the *coverage probability* is closer to the nominal value (95% e.g.)
- *Coverage probability* - proportion of the time that the interval contains the true value of interest
- Can include in addition *continuity correction* to ensure that the minimum coverage probability closer to the nominal value.
- You do not need to know the details! It's all done by `prop.test`

Wilson confidence intervals (second approximation)

```
> prop.test(2,30)

	1-sample proportions test with continuity correction

data:  2 out of 30, null probability 0.5
X-squared = 20.8333, df = 1, p-value = 5.01e-06
alternative hypothesis: true p is not equal to 0.5
95 percent confidence interval:
 0.01163184 0.23507287
sample estimates:
         p
0.0666666
```

Clopper-Pearson confidence intervals (“exact”)

“exact” 95% confidence interval

- It is based on the cumulative probabilities of the binomial distribution rather than an approximation.
- The interval may be wider than it needs to be to achieve X% confidence.
- Note, Wald & Wilson confidence intervals may be narrower than their nominal confidence width.
- It can be obtained on R using `binom.test`.

Clopper-Pearson confidence intervals ("exact")

```
> binom.test(2,30)

        Exact binomial test

data:  2 and 30
number of successes = 2, number of trials = 30,
  p-value = 8.68e-07
alternative hypothesis: true probability of success
  is not equal to 0.5
95 percent confidence interval:
 0.008178134 0.220735402
sample estimates:
probability of success
            0.06666667
```

Confidence intervals for p

Comparison of Confidence Intervals

Wald (approx.)	$(-0.022, 0.156)$
Wilson* (approx.)	$(0.012, 0.235)$
Clopper-Pearson ("exact")	$(0.008, 0.221)$

* Adjusted CI i.e. with continuity correction. Without continuity correction: (0.018,0.213)

More than two categories

Alcohol and nicotine consumption during pregnancy:
study of 452 mothers.

Nicotine (milligrams/day)

	None	1 – 15	16 or more	total
	304	65	83	452
proportion	0.673	0.144	0.184	

Proceed as before with each of the three proportions.

More than two categories - collapse categories

Alcohol and nicotine consumption during pregnancy: study of 452 mothers.

Nicotine (milligrams/day)

	None	1 – 15	16 or more	total
	304	65	83	452
proportion	0.673	0.144	0.184	

Nicotine (milligrams/day)

	None	1 or more	total
	304	$65 + 83$	452
proportion	0.673	$0.144 + 0.184$	

Proceed as before with difference of two proportions.

Difference between two proportions

If $Z_1 \stackrel{\mathrm{d}}{=} \mathrm{N}(\mu_1, \sigma_1)$, and $Z_2 \stackrel{\mathrm{d}}{=} \mathrm{N}(\mu_2, \sigma_2)$ are independent then

$$Z_1 - Z_2 \stackrel{\mathrm{d}}{=} \mathrm{N}\big(\mu_1 - \mu_2, \sqrt{\sigma_1^2 + \sigma_2^2}\,\big)$$

Therefore approximately,

$$\hat{P}_1 - \hat{P}_2 \stackrel{\mathrm{d}}{\approx} \mathrm{N}\left(p_1 - p_2, \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}\right)$$

Approximate 95% confidence interval for $p_1 - p_2$:

$$\hat{p}_1 - \hat{p}_2 \pm 1.96\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}.$$

Available on R, using `prop.test`.

Example: proportion of left-handed men and women

Random sample of 50 men: 5 left-handed. $5/50 = 0.1$.

Random sample of 30 women: 2 left-handed. $2/30 = 0.067$.

Approximate 95% confidence interval for $p_1 - p_2$:

$$0.1 - 0.067 \pm 1.96\sqrt{\frac{0.1\times 0.9}{50} + \frac{0.067\times 0.933}{30}}$$

$$= \ 0.033 \pm 0.122$$

$$= (-0.089, 0.156).$$

Example: proportion of left-handed men and women

```
> prop.test(x=c(5,2),n=c(50,30))     # correct=T is the default
  2-sample test ... with continuity correction
data:  c(5, 2) out of c(50, 30)
X-squared = 0.0104, df = 1, p-value = 0.9186
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.1153255  0.1819922
sample estimates:
    prop 1     prop 2
0.10000000 0.06666667
```

Approximate 95% confidence interval for $p_1 - p_2$:

$(-0.115, 0.182)$
recall: $(-0.089, 0.155)$ (no correction).