Statistical Methods

Lecture 8 – Estimation

Luke Dickens

Autumn 2020

Point estimators of central tendency

Estimating a population variance

Interval estimator of normal mean with known variance

Interval estimators of normal mean with unknown variance

Interval estimators of a population proportion

Point estimator of a population mean

If we have a sample $X_1, \ldots X_n$ from a population with unknown mean μ then we can estimate this mean with the sample mean, $\bar{X} = \left(\sum_i X_i\right)/n$ since:

$$E[\bar{X}] = \mu$$

We say that $\bar{X}$ is an **unbiased estimator** of μ.

Definition: An **estimator** is RV whose value is a statistic of a sample, and which predicts the value of a population parameter.

Definition: An **unbiased estimator** is one whose expected value equals the parameter it estimates.

How accurate is the sample mean?

We have seen that for a population with mean μ and SD σ, and a sample of size n, then

$$SD(\bar{X}) = \frac{\sigma}{\sqrt{n}}$$

and, for a large enough sample, $\bar{X}$ is approximately normal.

This tells us that:

- $\bar{X}$ is very likely to be in $[\mu - 2\sigma/\sqrt{n}, \mu + 2\sigma/\sqrt{n}]$
- A larger population SD leads to more variability (linearly proportional to σ)
- A larger sample leads to less variability (inversely proportional to $\sqrt{n}$)

Point estimator of a proportion

Let's say we are estimating the proportion, p, of a large population for which some characteristic is true. We take a sample of size n, and say X is the number in our sample with the characteristic. Then our estimator:

$$\hat{p} = \frac{X}{n}$$

Is a RV with:

$$E[\hat{p}] = p$$

$$SD(\hat{p}) = \sqrt{\frac{p(1-p)}{n}}$$

How accurate is $\hat{p}$?

Given $\hat{p}$ is an unbiased estimator with: $SD(\hat{p}) = \sqrt{\frac{p(1-p)}{n}}$

We also know that: $p(1-p) \leq \frac{1}{4}$

and so: $SD(\hat{p}) \leq \sqrt{\frac{1}{4n}} = \frac{1}{2\sqrt{n}}$

With large enough n (for normality), can say it is very likely that:

$$\hat{p} \in \left[p - \frac{1}{\sqrt{n}}, \quad p + \frac{1}{\sqrt{n}} \right]$$

This may be very conservative if p is close to 0 or 1.

Estimating the probability of a sensitive event

[Ros17, Sec. 8.3.1] presents an approach for estimating a proportion from study participants' responses to a yes/no question, where they may be reluctant to give one answer. Let's assume **yes** is the sensitive answer:

- Tell participants to secretly toss a coin:
  - if heads participant say **yes**
  - if tails they answer truthfully
- say true proportion of **yeses** is p and true proportion of **nos** is $q = (1 - p)$
- actual proportion of **nos** is $\frac{q}{2}$, so we can estimate this and hence q and p
- The *price* to pay is increased variance in the estimator

Point estimators of central tendency

Estimating a population variance

Interval estimator of normal mean with known variance

Interval estimators of normal mean with unknown variance

Interval estimators of a population proportion

Sample variance

With sample $X_1, \ldots, X_n$ from a population with unknown mean μ and unknown variance σ^2, we can estimate σ^2 with:

$$S^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}$$

Where

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

This is an **unbiased estimator**, i.e.

$$E[S^2] = \sigma^2$$

And we use $S = \sqrt{S^2}$ to estimate the population standard deviation σ.

Why divide by $n-1$ and not n?

If the mean μ was known then we have unbiased estimator with:

$$S^2 = \frac{\sum_{i=1}^{n}(X_i - \mu)^2}{n}$$

Since

$$E[S^2] = E[(X_i - \mu)^2] = \sigma^2$$

Intuition: $\bar{X}$ minimises the average squared distance to observed datapoints, and so average squared distance to μ is larger.

Point estimators of central tendency

Estimating a population variance

Interval estimator of normal mean with known variance

Interval estimators of normal mean with unknown variance

Interval estimators of a population proportion

Interval estimators

As point estimators are themselves RVs, we cannot expect them to be precise but we expect them to be *close* to the parameter they estimate.

An alternative to point estimators is to define an interval that very probably contains the parameter value.

Definition: An interval estimator is an interval that is predicted to contain the population parameter it seeks to estimate. The associated confidence is the probability with which we expect the value to lie in that interval.

We can do this by using the probability distribution of the point estimator.

Estimating a 95% confidence interval

Assume we have a normal distribution with known SD σ and unknown mean μ, and a sample $X_1, \ldots, X_n$ of size n.

The sample mean $\bar{X}$ is a normally distributed RV with mean μ and SD $\sigma/\sqrt{n}$ and hence we can define the standard normal RV

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} = \frac{\sqrt{n}}{\sigma}\left(\bar{X} - \mu\right)$$

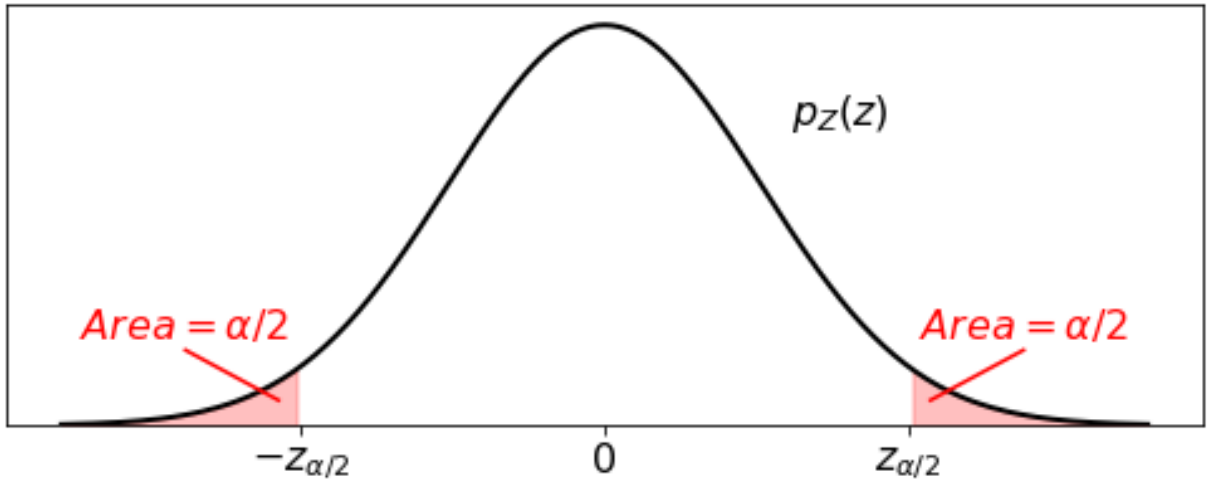
Consider that Z lies between $z_{0.975} = -1.96$ and $z_{0.025} = 1.96$ with a 95% probability and so:

$$\Pr\left(\frac{\sqrt{n}}{\sigma}|\bar{X} - \mu| < 1.96\right) = 0.95$$

$$= \Pr\left(\bar{X} - 1.96\frac{\sqrt{n}}{\sigma} < \mu < \bar{X} + 1.96\frac{\sqrt{n}}{\sigma}\right)$$

And so $\bar{X} \pm 1.96\frac{\sqrt{n}}{\sigma}$ is a 95% confidence interval for μ.

Other confidence intervals



The percentiles of the standard normal (combined with symmetry) help us to define confidence intervals.

Confidence level $100(1-\alpha)$	**Corresponding value of** α	**Value of** $z_{\alpha/2}$
90	0.1	$z_{0.05} = 1.645$
95	0.05	$z_{0.025} = 1.960$
99	0.01	$z_{0.005} = 2.576$

Choosing n

To obtain a $100(1-\alpha)$ confidence interval of a prespecified length, we can work in reverse to determine the required size of sample.

We know that a sample of size n will lead to a $100(1-\alpha)$ confidence interval in the range

$$\bar{X} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

Suppose we want an interval of no more than length b then:

$$2z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq b$$

Rearranging gives:

$$n \geq \left(\frac{2\, z_{\alpha/2}\, \sigma}{b}\right)^2$$

Lower and upper confidence bounds

We may want to state that with probability P a mean is above (below) a given lower (upper) bound.

By a similar argument as before, we define a standard normal RV

$$Z = \sqrt{n}\frac{\bar{X} - \mu}{\sigma}$$

and say:

$$\Pr\left(\sqrt{n}\frac{\bar{X} - \mu}{\sigma} < z_\alpha\right) = 1 - \alpha$$

$$= \Pr\left(\mu > \bar{X} - z_\alpha \frac{\sigma}{\sqrt{n}}\right)$$

In fact: For large enough n, these interval, sample size and bound estimates hold irrespective of the underlying distribution.

Point estimators of central tendency

Estimating a population variance

Interval estimator of normal mean with known variance

Interval estimators of normal mean with unknown variance

Interval estimators of a population proportion

A standardised RV for unknown variance

Consider a normal population with unknown mean μ and **unknown SD** σ. We have a sample $X_1, \ldots, X_n$ of size n and want to estimate μ.

If we knew σ, then we could obtain a standard normal RV with

$$Z = \sqrt{n}\frac{\bar{X} - \mu}{\sigma}$$

Replacing σ with sample standard deviation estimate S gives:

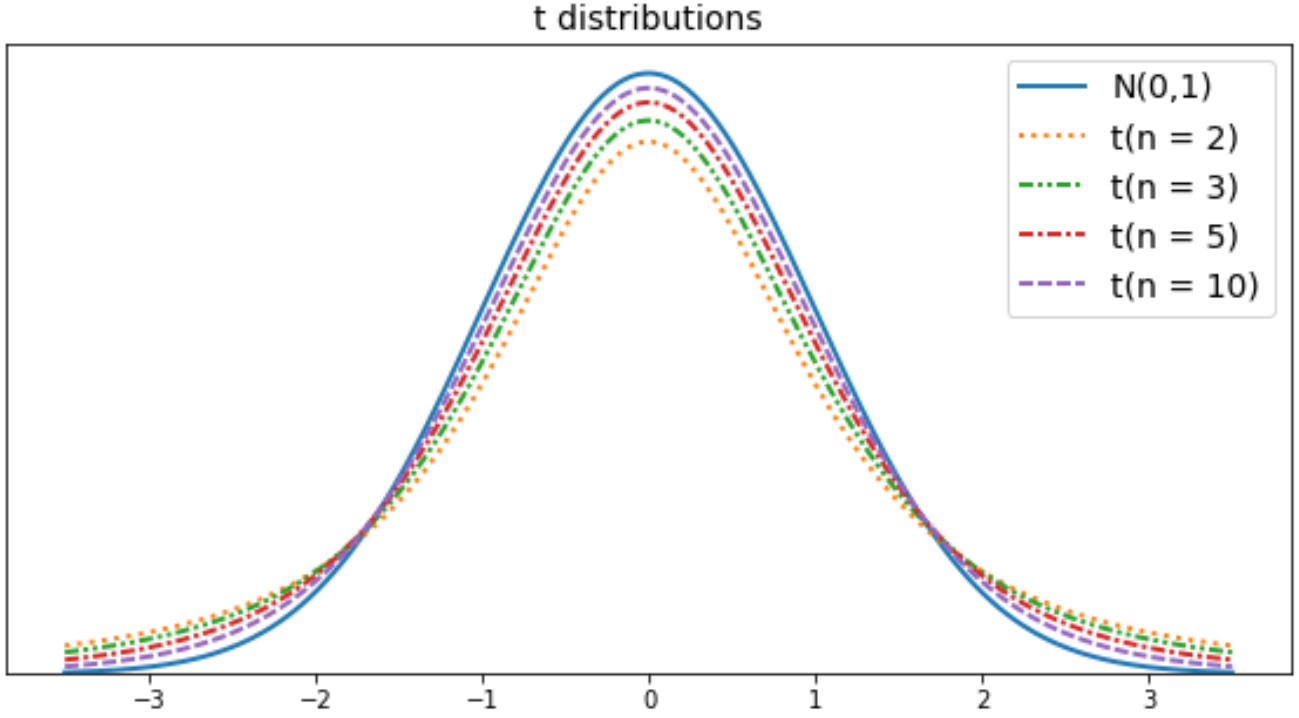
$$T_{n-1} = \sqrt{n}\frac{\bar{X} - \mu}{S}$$

T_{n-1} is not a standard normal RV but from t closely related (Student's) t distribution with $n - 1$ degrees of freedom.

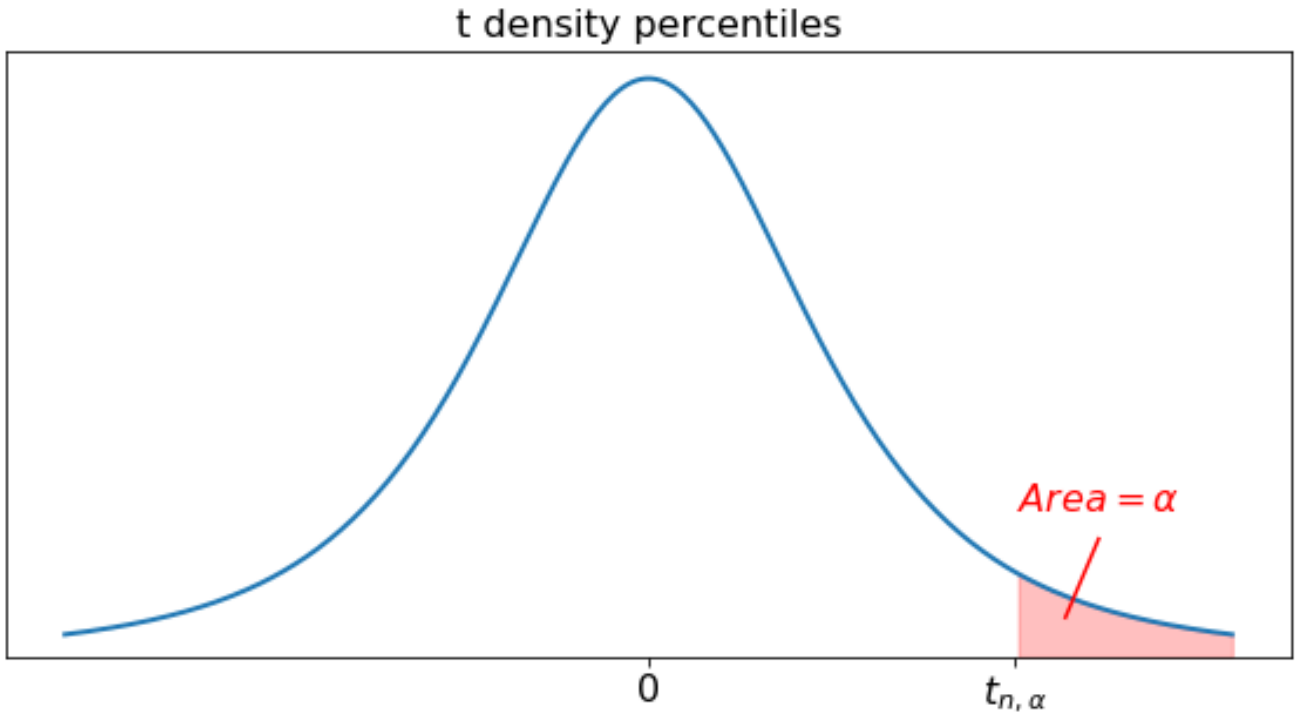
Student's t distribution and T_{n-1}

We say T_{n-1} has $n-1$ degrees of freedom since $(n-1)S^2/\sigma^2$ is a chi-squared RV with $n-1$ degrees of freedom.

The probability density of a t distribution looks similar to a standard normal (increasingly so as n gets larger). For small n the tails are significantly heavier.



t density percentiles



The t-density percentile: $\Pr(T_n > t_{n,\alpha}) = \alpha$

Just like the standard normal distribution, we can calculate percentiles of t-distributions, $t_{n,\alpha}$ (see [Ros17, Tbl. D.2] for common values), and use these to define intervals or lower bounds on our estimates for μ.

Predicting intervals with the t-distribution

We can predict interval estimators for μ even when the SD σ is approximated with sample SD S. Recall:

$$T_{n-1} = \sqrt{n}\frac{\bar{X} - \mu}{S}$$

For $100(1-\alpha)\%$ confidence we say:

$$\Pr\left(\sqrt{n}\frac{\left|\bar{X} - \mu\right|}{S} < t_{n-1,\alpha/2}\right) = 1 - \alpha$$

$$= \Pr\left(\bar{X} - t_{n-1,\alpha/2}\frac{S}{\sqrt{n}} < \mu < \bar{X} + t_{n-1,\alpha/2}\frac{S}{\sqrt{n}}\right)$$

And so a $100(1-\alpha)\%$ confidence interval is given by:
$\bar{X} \pm t_{n-1,\alpha/2}\frac{S}{\sqrt{n}}$

Upper and lower bounds with unknown SD

We can predict lower or upper bounds too. Recall that we have a size n sample, from a normal population with mean μ and SD σ:

A $100(1-\alpha)\%$ confidence lower bound for μ is given by:

$$\bar{X} - t_{n-1,\alpha}\frac{S}{\sqrt{n}}$$

A $100(1-\alpha)\%$ confidence upper bound for μ is given by:

$$\bar{X} + t_{n-1,\alpha}\frac{S}{\sqrt{n}}$$

Note we are using $t_{n-1,\alpha}$ (**not** $t_{n-1,\alpha/2}$ as with intervals).

Point estimators of central tendency

Estimating a population variance

Interval estimator of normal mean with known variance

Interval estimators of normal mean with unknown variance

Interval estimators of a population proportion

Recalling proportion point estimator $\hat{p}$

Recall: If we have a population a proportion p of whom have some characteristic, and we randomly sample n elements and determine that X have the characteristic, then $\hat{p} = X/n$ is an unbiased estimator of p and

$$E[\hat{p}] = p$$

$$SD(\hat{p}) = \sqrt{\frac{p(1-p)}{n}}$$

Moreover: If n is large enough such that $np > 5$ and $n(1-p) > 5$ then $\hat{p}$ is approximately normal.

Interval estimator for $\hat{p}$

Given the properties on the last slide, a $100(1-\alpha)\%$ confidence interval estimator of p is given by

$$\hat{p} \pm z_{\alpha/2} SD(\hat{p})$$

but $SD(\hat{p})$ is given in terms of p which is unknown.

Instead, we can approximate the SD with

$$SD(\hat{p}) \approx \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

And we can calculate the interval as:

$$\hat{p} \pm z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Length of confidence interval for $\hat{p}$

We have a $100(1-\alpha)\%$ interval estimator for a proportion given by

$$p \in \left[\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \quad \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \right]$$

The length of this interval is

$$2z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

where $\hat{p}$ is the sample proportion and $\hat{p}(1-\hat{p}) \leq \frac{1}{4}$ so

$$\text{interval length} \ \leq \ \frac{z_{\alpha/2}}{\sqrt{n}}$$

Choosing n for interval $\hat{p}$

We can now estimate a sample size required to obtain a $100(1-\alpha)\%$ confidence interval for p whose length is less than some prespecified value, say b, as any value n such that

$$\frac{z_{\alpha/2}}{\sqrt{n}} < b$$

Rearranging this inequality and squaring both sides gives

$$n > \left(\frac{z_{\alpha/2}}{b}\right)^2$$

And any n that satisfies this inequality will suffice.

Lower and upper confidence bounds on $\hat{p}$

Again, we can consider only one bound on our estimator.

A $100(1-\alpha)\%$ lower confidence bound on p is given by

$$\hat{p} - z_\alpha \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

A $100(1-\alpha)\%$ upper confidence bound on p is given by

$$\hat{p} + z_\alpha \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

References I

[Ros17] Sheldon M. Ross, *Introductory Statistics*, 4 ed., Academic Press, 2017.