Statistical Methods: Lecture 5

Lecture Overview

Sampling distributions and estimators

Estimating a Population Mean

Estimating a Population Proportion

5.4 Sampling distributions and estimators

Example: file sizes

A statistics teacher has 2692 files related to Statistical Methods. What is the average file size (population mean) μ?

Take (representative) sample of size n from population.
Compute $\overline{x}_n$ and use as estimate of μ. Is it good?

Example: Brexit

UK's referendum on June 23, 2016: stay in or leave EU?
Ca. 46.5 million Britons could vote "remain" or "leave".
Population proportion p denotes proportion of Britons that votes "remain".

3 days before referendum, Survation conducted a poll: excluding undecided, out of $n = 893$ Britons, 50.6% would vote "remain".

Sample proportion $\hat{p}_{893} = 0.506$. Is it a good estimate of population proportion p?

What if we selected some other n files, or asked some other 893 Britons?

5.4 Sampling distributions and estimators

We cannot say whether $\overline{x}_n$ is close to μ, or whether $\hat{p}_n$ is close to p, but we can study the distribution of all possible values of $\overline{X}_n$ or $\hat{P}_n$ for fixed sample size n.

Definition: Sampling distribution of the sample mean

Let the random variable $\overline{X}_n$ denote the sample mean of a sample of size n.
The sampling distribution of the sample mean consists of all possible values of $\overline{X}_n$, based on all possible samples of size n, and corresponding probabilities.

You don't want to compute this for $n > 2$. Luckily:

The Central Limit Theorem (CLT)

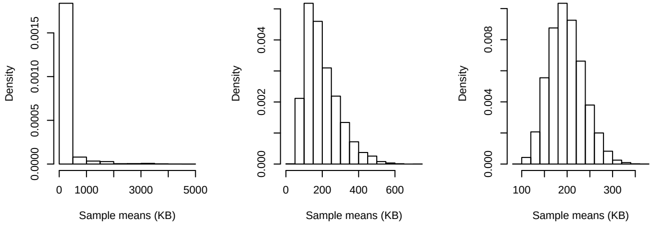
Independently draw a sample of size $n > 30$ from a population with mean μ and standard deviation σ. Then $\overline{X}_n$ has approximately a $N(\mu, \frac{\sigma^2}{n})$-distribution.

Sampling distribution (of random variable) $\neq$ sample distribution (of dataset).

5.4 Sampling distributions and estimators

Example: file sizes – approximating sampling distribution

File sizes of 2692 teaching files.
Left: distribution of 10 000 values of sample mean (i.e. approximation of sampling distribution); sample size $n = 5$.
Middle: $n = 100$, right: $n = 500$.



5.4 Sampling distributions and estimators

Sampling distribution of sample proportion

This is the probability distribution of random variable $\hat{P}_n$: consists of all possible values of $\hat{p}_n$ based on all possible samples of size n and corresponding probabilities.

Sample proportion: special case of sample mean!
Population proportion p (i.e. prob. of "remain").
Individual answers: realizations of random variables X_i with values 1/0 (yes/no);
$P(X=1)=p$ and $P(X=0)=1-p$, where $p=$ population proportion.

If n people surveyed, we get $x_1, x_2, \ldots, x_n$:

$$x_i = \begin{cases} 1 & \text{if subject } i \text{ said 'yes' } / \text{ has the property} \\ 0 & \text{if subject } i \text{ said 'no' } / \text{ does not have the property} \end{cases}$$

Then $\hat{p}_n = (x_1 + x_2 + \ldots + x_n)/n$.

5.4 Sampling distributions and estimators

Finding sampling distribution of sample proportion

Recall $P(X = 1) = p$, $P(X = 0) = 1 - p$.
Use CLT... need population mean and population standard deviation:

$$\mu = 1 \cdot p + 0 \cdot (1 - p) = p$$
$$\sigma = \sqrt{p(1 - p)}$$

Sampling distribution for large n

For large n (> 30) the sample proportion $\hat{P}_n$ of a population with population proportion p is approximately normal with mean p and standard deviation $\sqrt{p(1 - p)/n}$, i.e., approximately

$$\hat{P}_n \sim N\Big(p, \frac{p(1 - p)}{n}\Big).$$

5.4 Sampling distributions and estimators: recap

Population mean: μ; population standard deviation: σ

Sampling distribution of sample mean

For large n (> 30), the sampling distribution of $\overline{X}_n$ is approximately normal with mean μ and standard deviation $\sigma/\sqrt{n}$.

Population proportion: p

Sampling distribution of sample proportion

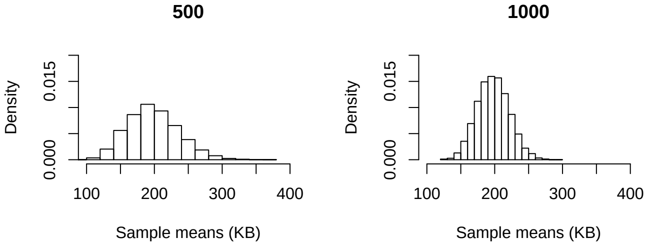
For large n (> 30), the sampling distribution of $\hat{P}_n$ is approximately normal with mean p and standard deviation $\sqrt{p(1-p)/n}$.

6.3 Estimating a Population Mean

Sampling distribution of sample mean (estimator) approximately normally distributed.
For a given sample, the estimator yields an estimate of population mean μ. Accuracy?

For any n, unbiased: $E(\overline{X}_n) = \mu$ ("targets the population mean μ").
For one sample, we obtain only one estimate.
Standard deviation of the sampling distribution: how good is the estimate.

Recall the "files" example:



6.3 Estimating a Population Mean

Use approximate distribution to construct confidence intervals:

95% confidence interval for μ:

range of estimator values; we are 95% confident that this interval actually contains μ.

"95% confident..."

For 100 independent samples of size n, calculate confidence intervals for each.
On average, 95 of them contain μ.

Incorrect intepretation

For a given 95% confidence interval it does not mean: 95% chance that μ is in this interval, μ is fixed and unknown, interval is a realization of a random interval.

6.3 Estimating a Population Mean

Recall: $\overline{X}_n \sim N(\mu, \sigma^2/n)$ (approx.).

If σ unknown: $\overline{X}_n \sim N(\mu, s_n^2/n)$ approx.
Here, s_n = sample standard deviation.

Recall: $Z = \dfrac{\overline{X}_n - \mu}{s_n/\sqrt{n}} \sim N(0,1)$ (approx.) and use Table 2:

$$0.95 = P(-1.96 \leq Z \leq 1.96) = P\left(\mu - 1.96\frac{s_n}{\sqrt{n}} \leq \overline{X}_n \leq \mu + 1.96\frac{s_n}{\sqrt{n}}\right)$$

Exactly what we need. Why?
Because for (approximately) 95 out of 100 independent samples of size n

$$\mu - 1.96\frac{s_n}{\sqrt{n}} \leq \overline{x}_n \leq \mu + 1.96\frac{s_n}{\sqrt{n}}$$

which is equivalent to

$$\overline{x}_n - 1.96\frac{s_n}{\sqrt{n}} \leq \mu \leq \overline{x}_n + 1.96\frac{s_n}{\sqrt{n}}$$

6.3 Estimating a Population Mean

Definition: 95% confidence interval (CI) for μ

$E = 1.96\frac{s_n}{\sqrt{n}}$ is called the margin of error, and the interval

$$\left[\overline{x}_n - 1.96\frac{s_n}{\sqrt{n}}, \overline{x}_n + 1.96\frac{s_n}{\sqrt{n}}\right]$$

is called a 95% confidence interval for μ. (If σ is known, use it instead of s_n)

Example: program files

Randomly selected $n = 144$ files with $\overline{x}_n = 150.53$ and $s_n = 502.75$.
95% confidence interval for μ given by

$$\left[150.53 - 1.96\frac{502.75}{\sqrt{144}}, 150.53 + 1.96\frac{502.75}{\sqrt{144}}\right] = [68.41, 232.65]$$

Interpretation

Don't know whether true μ is in this particular CI or not.
If we constructed 100 confidence intervals based on 100 independent samples of size 144, approximately 95 of them would contain μ.

6.3 Estimating a Population Mean

Slightly different than in the book

- Book uses *t-distribution* to construct CI's. We will do that later.
- $z_{\alpha/2} = 1.96$ for $\alpha = 0.05 = 1 - 0.95$,
  $z_{\alpha/2} = z$ score separating an area of $\alpha/2$ in the right tail of $N(0,1)$:

  $$P(Z \geq z_{\alpha/2}) = \alpha/2 \quad \text{and} \quad P(Z \leq -z_{\alpha/2}) = \alpha/2$$

  so by properties of probability

  $$P(-z_{\alpha/2} \leq Z \leq z_{\alpha/2}) = 1 - \alpha$$

- Sometimes 2 is used instead of 1.96 – see rule of thumb for $N(0,1)$.

6.3 Estimating a Population Mean

Margin of error $E = 1.96\frac{s_n}{\sqrt{n}}$. Choose n so that E as small as desired:

First, fix an estimate of standard deviation.
E.g., sample standard deviation (or Range/4).
Let us denote it by σ. Then

$$E = 1.96\frac{s_n}{\sqrt{n}} \approx 1.96\frac{\sigma}{\sqrt{n}} \leq E_{max} \quad \Leftrightarrow \quad n \geq \left(\frac{1.96 \cdot \sigma}{E_{max}}\right)^2$$

6.2 Estimating a Population Proportion

Very similar to population mean, hence this part is more brief

Recall: $\hat{P}_n \sim N(p, p(1-p)/n)$ (approx.).

Again estimate standard deviation: $\hat{P}_n \sim N(p, \hat{p}_n(1-\hat{p}_n)/n)$ (approx.).

Definition: 95% confidence interval (CI) for p

$E = 1.96\sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n}}$ is called the margin of error, and the interval

$$\left[\hat{p}_n - 1.96\sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n}}, \hat{p}_n + 1.96\sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n}}\right]$$

is called a 95% confidence interval for p.

6.2 Estimating a Population Proportion

Example: Brexit

Based on answers of $n = 893$ Britons: sample proportion $\hat{p}_{893} = 0.506$.

95% confidence interval for p:

$$\left[0.506 - 1.96\sqrt{\frac{0.506 \cdot 0.494}{893}}, 0.506 + 1.96\sqrt{\frac{0.506 \cdot 0.494}{893}}\right] = [0.473, 0.539]$$

Interpretation?

6.2 Estimating a Population Proportion

Margin of error $E = 1.96\sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n}}$. Choose n so that E as small as we want.

Population proportion is always between 0 and 1, so $p(1-p) \leq 0.25$. Then

$$E = 1.96\sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n}} \leq 1.96\sqrt{\frac{1}{4n}} \leq E_{max} \quad \Leftrightarrow \quad n \geq \left(\frac{1.96}{2 \cdot E_{max}}\right)^2$$

This bound can be too conservative if the true p is far from 0.5. Alternatives?

Other percentages

If $Z \sim N(0,1)$,

$$P(-1.96 \leq Z \leq 1.96) = 0.95;$$

other standard normal quantiles $\rightsquigarrow$ other confidence levels.

90% confidence

$$P(-1.645 \leq Z \leq 1.645) = 0.9$$

Margins of errors are

$$1.645 \frac{s_n}{\sqrt{n}} \qquad \text{and} \qquad 1.645 \sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n}}$$

99% confidence

$$P(-2.575 \leq Z \leq 2.575) = 0.99$$

Margins of errors are

$$2.575 \frac{s_n}{\sqrt{n}} \qquad \text{and} \qquad 2.575 \sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n}}$$

Estimating Population Mean and Population Proportion: recap

Population mean

Sample mean is used to estimate population mean.
95% confidence interval is given by $\left[\overline{x}_n - 1.96\frac{s_n}{\sqrt{n}}, \overline{x}_n + 1.96\frac{s_n}{\sqrt{n}}\right]$
For the margin of error $E = 1.96\frac{s_n}{\sqrt{n}}$ to be smaller than E_{max} we need sample size $n \geq \left(\frac{1.96 \cdot \sigma}{E_{max}}\right)^2$

Population proportion

Sample proportion is used to estimate population proportion.
95% confidence interval is given by $\left[\hat{p}_n - 1.96\sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n}}, \hat{p}_n + 1.96\sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n}}\right]$
For the margin of error $E = 1.96\sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n}}$ to be smaller than E_{max} we need sample size $n \geq \left(\frac{1.96}{2 \cdot E_{max}}\right)^2$