DSA 8010 - foundations of inference

Inferential statistics

Review: populations and samples

Population. Entire set of measurements of interest. Parameters are numeric characteristics of a population and are often unknown.

Sample. Subset of measurements from the population of interest. Samples contain information that is limited and variable. Statistics are calculated from observed sample values.

Goal in collecting data: ensure that data are representative of a broader population.

Types of statistical analyses

Descriptive. Summarize a data set.

Inferential. Draw conclusions about the population using sample data.

Example. A researcher wants to learn about the sizes of a certain species of trout in Western NC. He (somehow) takes a simple random sample of $n = 12$ trout from the population and records their lengths.

Types of statistical analyses

Descriptive. Summarize a data set.

Inferential. Draw conclusions about the population using sample data.

Example. A researcher wants to learn about the sizes of a certain species of trout in Western NC. He (somehow) takes a simple random sample of $n = 12$ trout from the population and records their lengths.

The average of his sample of trout was 7 cm. What does that tell us about the scientific question?

Types of statistical analyses

Descriptive. Summarize a data set.

Inferential. Draw conclusions about the population using sample data.

Example. A researcher wants to learn about the sizes of a certain species of trout in Western NC. He (somehow) takes a simple random sample of $n = 12$ trout from the population and records their lengths.

The average of his sample of trout was 7 cm. What does that tell us about the scientific question?

What do we know about μ, the population average length?

Example (clinical trials)

A clinical trial randomly assigned n_1 patients to receive a placebo and n_2 patients to receive an experimental drug believed to reduce blood pressure. The response variable is the systolic blood pressure and was measured on each patient.

The placebo group had a mean systolic of $\bar{y}_1 = 113.4$. The drug group had a mean systolic of $\bar{y}_2 = 105.7$.

Is the drug effective?

Example (clinical trials)

A clinical trial randomly assigned n_1 patients to receive a placebo and n_2 patients to receive an experimental drug believed to reduce blood pressure. The response variable is the systolic blood pressure and was measured on each patient.

The placebo group had a mean systolic of $\bar{y}_1 = 113.4$. The drug group had a mean systolic of $\bar{y}_2 = 105.7$.

Is the drug effective?

Is $\mu_2 < \mu_1$?

Parameters and statistics

- parameter: numerical characteristic of a population. Denoted using Greek letters (e.g. μ =population mean.)
- statistic: numerical characteristic of a sample. Typically denoted using Latin letters (e.g. $\bar{y}$ = sample mean.)

Parameters that we often want to know:

- μ, the mean or expected value of a population.
- σ, the standard deviation of a population.
- π, the probability of success or proportion of successes in a population.
- $\mu_1 - \mu_2$, the difference between the means of two groups in a population.

Inference overview

Inference

We generally divide inferential procedures into two categories:

Estimation, in which we provide an estimate of a parameter along with appropriate uncertainty quantification; and

Testing, in which we posit a value of a parameter and consider whether the data provide sufficient evidence to negate that value.

We will see, however, that testing and estimation are tied closely together and are not all that different.

Example: approval ratings

Adapted from *Open Intro Statistics, 4th edition, Diez et al*. A recent poll of 25,000 registered voters found that 11,250 approved of the US President's performance, suggesting an approval rating of 45%.

Some inferential follow-ups:

Estimation. Find a confidence interval for π, the true proportion of all registered voters who approve of the president's performance.

Testing. Do the data provide evidence that the approval rating is significantly different than that of his predecessor, 0.43%?

Example: approval ratings

Adapted from *Open Intro Statistics, 4th edition, Diez et al*. A recent poll of 25,000 registered voters found that 11,250 approved of the US President's performance, suggesting an approval rating of 45%.

Some inferential follow-ups:

Estimation. Find a confidence interval for π, the true proportion of all registered voters who approve of the president's performance.

```
> prop.test(11250, 25000, conf.level=.95)

        1-sample proportions test with continuity
        correction

data:  11250 out of 25000, null probability 0.5
X-squared = 249.8, df = 1, p-value < 2.2e-16
alternative hypothesis: true p is not equal to 0.5
95 percent confidence interval:
 0.4438213 0.4561941
```

Interpretation. We are 95% confident that the true proportion of all registered voters who approve of the president's performance is between 44.38% and 45.62%.

Example: approval ratings

Adapted from *Open Intro Statistics, 4th edition, Diez et al*. A recent poll of 25,000 registered voters found that 11,250 approved of the US President's performance, suggesting an approval rating of 45%.

Some inferential follow-ups:

Estimation. Find a confidence interval for π, the true proportion of all registered voters who approve of the president's performance.

Testing. Do the data provide evidence that the approval rating is significantly different than that of his predecessor, 0.43%?

Example: approval ratings

Adapted from *Open Intro Statistics, 4th edition, Diez et al*. A recent poll of 25,000 registered voters found that 11,250 approved of the US President's performance, suggesting an approval rating of 45%.

Some inferential follow-ups:

Testing. Do the data provide evidence that the approval rating is significantly different than that of his predecessor, 0.43%?

```
> prop.test(11250, 25000, p=.43)

        1-sample proportions test with continuity
        correction

data:  11250 out of 25000, null probability 0.43
X-squared = 40.718, df = 1, p-value = 1.758e-10
alternative hypothesis: true p is not equal to 0.43
```

Standard errors

Standard error

Consider a random sample of $n = 50$ Bernoulli trails with $\pi = 0.85$.

If we observed $X = 42$ successes, the sample proportion is $\widehat{\pi} = 42/50 = 0.84$.

Standard error

Consider a random sample of $n = 50$ Bernoulli trails with $\pi = 0.85$.

If we observed $X = 42$ successes, the sample proportion is $\widehat{\pi} = 42/50 = 0.84$.

Due to random variability, we might observe 40, 46, or even 30 successes, giving us different values of the sample proportion.

Standard error

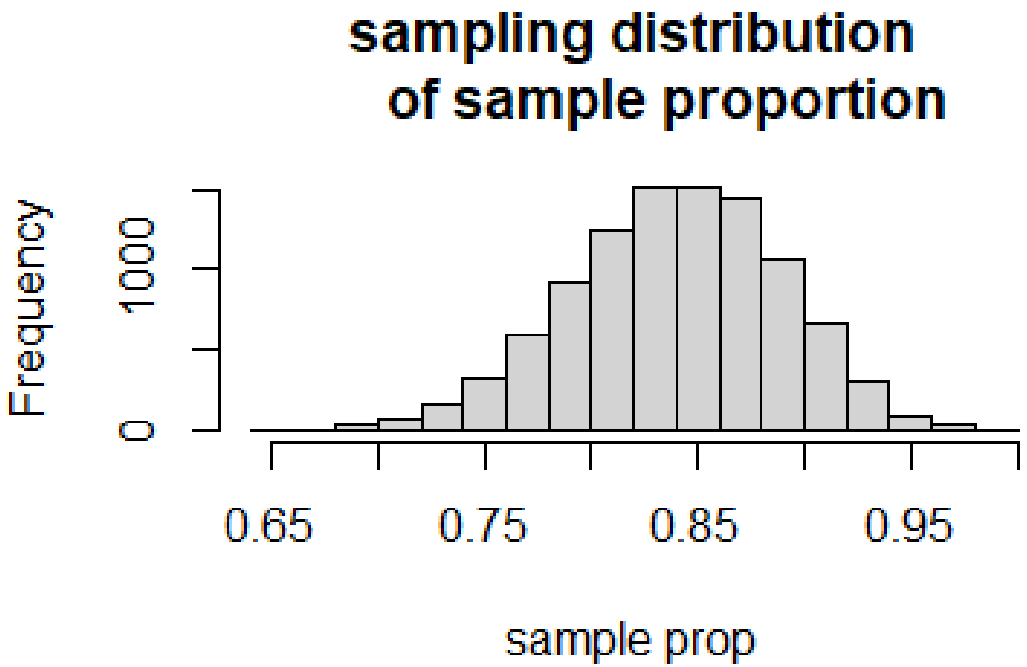
Consider a random sample of $n = 50$ Bernoulli trails with $\pi = 0.85$.

If we observed $X = 42$ successes, the sample proportion is $\widehat{\pi} = 42/50 = 0.84$.

Due to random variability, we might observe 40, 46, or even 30 successes, giving us different values of the sample proportion.

```
> # generate a binomial random variable
> n <- 50
> x<- rbinom(1,size=n,prob=0.85)
> x
[1] 44
> # calculate the sample proportion
> x/n
[1] 0.88
> # generate a binomial random variable
> n <- 50
> x<- rbinom(1,size=n,prob=0.85)
> x
[1] 41
> # calculate the sample proportion
> x/n
[1] 0.82
```

Standard error



A standard error measures how variable a statistic is in repeated sampling.

Standard error

- A standard error describes the variability of the statistics that could be calculated from a sample of size n.
- Standard errors are typically not observed directly because they arise from taking *repeated samples of size n* from the same target population.
- We typically have fixed formulas for standard errors. They usually depend on n.

Standard error

- A standard error describes the variability of the statistics that could be calculated from a sample of size n.
- Standard errors are typically not observed directly because they arise from taking *repeated samples of size n* from the same target population.
- We typically have fixed formulas for standard errors. They usually depend on n.

Example: The standard error of the sample proportion is $\sqrt{\frac{\pi(1-\pi)}{n}}$.

Estimation

Estimation

Estimation in a statistical sense refers to using data to obtain an educated guess of a population parameters.

Parameter that we often want to estimate:

- μ, the mean or expected value of a population.
- σ, the standard deviation of a population.
- π, the probability of success or proportion of successes in a population.
- $\mu_1 - \mu_2$, the difference between the means of two groups in a population.

Estimation

A point estimate is a statistic that gives a good guess of the population parameter.

Examples:

The sample mean is a point estimate of μ.

The sample proportion is a point estimate of π.

Estimation

A confidence interval is a range of plausible values that of the population parameter.

- The general form of a confidence interval is

$$\text{point estimate} \pm \text{margin of error}$$

- The point estimate is a good guess, but it is obtained from a sample that, due to chance alone, will not be exactly identical to the population. Confidence intervals account for this uncertainty.
- Confidence intervals take into account uncertainty due to both the sample size and the random variability of the data.
- The confidence level is a percentage that describes the accuracy of the interval in repeated sampling. 95% is often used.

Confidence interval

The structure of a confidence interval:

$$\text{point estimate} \pm \text{margin of error}$$

$$\text{point estimate} \pm \text{multiplier} * \text{standard error}$$

- The lower endpoint of the confidence interval has the form point estimate - multiplier times standard error.
- The upper endpoint of the confidence interval has the form point estimate + multiplier times standard error.
- The "multiplier times standard error" portion gives an appropriate amount of wiggle room away from the point estimate. It is referred to as the margin of error.

Confidence interval

A confidence interval has the following structure.

$$\text{point estimate} \pm \text{multiplier} * \text{standard error}$$

- The multiplier comes from the sampling distribution of the point estimate. It ensures that we have the specified confidence level.

  Often, this will be a normal distribution or a t distribution.
- The standard error is the standard deviation of the point estimate. It is smaller when the sample size is large.

Example confidence intervals

95% confidence interval for π.

$$\widehat{\pi} \pm z^*_{0.025} \sqrt{\frac{\widehat{\pi}(1-\widehat{\pi})}{n}}$$

95% confidence interval for μ.

$$\bar{y} \pm t^*_{n-1,0.025} \cdot s/\sqrt{n}$$

Testing

Testing: basic idea

- We start with a question about some population parameter.
- Turn the question into hypotheses, which are opposite statements about the value of a parameter.

  The alternative or research hypothesis is what we are hoping to support with the data.
  The null hypothesis is what we are hoping our data provide evidence against.
- Test statistics and p-values are used to quantify the degree of evidence against the null.
- We reject the null hypothesis if the p-value is small. This means that the data we observed are very unlikely to occur under the specified model and hypothesized parameter value.

Test for a proportion

Testing. Do the data provide evidence that the approval rating is significantly different than that of his predecessor, 43%?

```
> prop.test(11250, 25000, p=.43)

        1-sample proportions test with continuity
        correction

data:  11250 out of 25000, null probability 0.43
X-squared = 40.718, df = 1, p-value = 1.758e-10
alternative hypothesis: true p is not equal to 0.43
```

Test for a proportion

Testing. Do the data provide evidence that the approval rating is significantly different than that of his predecessor, 43%?

Hypotheses:

Test for a proportion

Testing. Do the data provide evidence that the approval rating is significantly different than that of his predecessor, 43%?

Test statistic:

Test for a proportion

Testing. Do the data provide evidence that the approval rating is significantly different than that of his predecessor, 43%?

P-value:

Test for a proportion

Testing. Do the data provide evidence that the approval rating is significantly different than that of his predecessor, 43%?

Decision:

Models

Statistical models

A statistical model represents observable data using a probability distribution.

- Statistical models allow us to calculate probabilities of observing data points or statistics and to make predictions about new data.
- The uncertainty in observable outcomes can then be quantified using probability.

Example (trout)

Example. A researcher wants to learn about the sizes of a certain species of trout in Western NC. He (somehow) takes a simple random sample of $n = 12$ trout from the population and records their lengths.

- The data are *numeric* measurements.
- One possible statistical model would assume that the observed lengths come from a Normal distribution.
- This is a good model if it is reasonable to think that the distribution of all WNC trout lengths is bell-shaped and symmetric.

Statistical models

A statistical model represents observable data using a probability distribution.

Many statistical models can be written in this format:

$$\text{data} = \text{fixed value(s)} + \text{noise}$$

where

- the noise has a Normal or other probability distribution; and
- the fixed portion is represented using parameters.

Example (trout)

Example. A researcher wants to learn about the sizes of a certain species of trout in Western NC. He (somehow) takes a simple random sample of $n = 12$ trout from the population and records their lengths.

Write a statistical model for the measurements.

Example (clinical trials)

A clinical trial randomly assigned n_1 patients to receive a placebo and n_2 patients to receive an experimental drug believed to lower blood pressure. The response variable is the systolic blood pressure and was measured on each patient.

The data might look like this:

	systolic	group
1	101	1
2	92	1
3	117	2
4	116	1
5	140	2
6	85	2
7	99	1
8	88	2
9	122	2
10	108	2

Example (clinical trials)

A clinical trial randomly assigned n_1 patients to receive a placebo and n_2 patients to receive an experimental drug believed to lower blood pressure. The response variable is the systolic blood pressure and was measured on each patient.

- Notation. Let Group 1 be the placebo group and Group 2 be the drug group. Denote the ith observation from Group 1 using Y_{i1}, $i = 1, \ldots, n_1$ and the ith observation from Group 2 using Y_{i2}, $i = 1, \ldots, n_2$.

Possible statistical model.

$$Y_{i1} = \mu_1 + \epsilon_{i1};$$
$$Y_{i2} = \mu_2 + \epsilon_{i2},$$

where the ϵ_{ij} are i.i.d. $N(0, \sigma^2)$ for all i and $j = 1, 2$.

Example (clinical trials)

A possible statistical model for these data would be the normal, two means model.

$$Y_{i1} = \mu_1 + \epsilon_{i1};$$
$$Y_{i2} = \mu_2 + \epsilon_{i2},$$

where the ϵ_{ij} are i.i.d. $N(0, \sigma^2)$ for all i and $j = 1, 2$.

Example (help-seeking)

A survey was administered to a random sample of students asking whether they would seek help from a parent if they were experiencing violence in a dating relationship.

The variable `parent` is equal to 1 if the student responded 'yes' and 0 if the student responded 'no'. Write a statistical model for this variable.

	A	B	C	
1	Wave	SID	parent	pe
2	1	134	1	
3	1	137	2	
4	1	139	2	
5	1	145	2	
6	1	146	2	
7	1	147	2	
8	1	148	1	
9	1	149	2	

Some statistical models

Typically with numeric data, we use use Normal models:

Normal, one mean. $Y_i = \mu + \epsilon_i$ where $\epsilon_1, \ldots, \epsilon_n$ are i.i.d. $N(0, \sigma^2)$.

Note: this is the same as saying $Y_1, \ldots, Y_n$ are i.i.d. $\mathrm{N}(\mu, \sigma^2)$.

Normal, two means. Denote the ith observation from Group 1 using Y_{i1}, $i = 1, \ldots, n_1$ and the ith observation from Group 2 using Y_{i2}, $i = 1, \ldots, n_2$.

Model:

$$Y_{i1} = \mu_1 + \epsilon_{i1};$$
$$Y_{i2} = \mu_2 + \epsilon_{i2},$$

where the ϵ_{ij} are i.i.d. $N(0, \sigma^2)$ for all i and $j = 1, 2$.

When we have binary data, we will typically use a Bernoulli/Binomial model:

Bernoulli model. $Y_1, \ldots, Y_n$ are i.i.d. with a Bernoulli(π) distribution.

I.I.D. assumption

The i.i.d. assumption

The assumption of independent and identically distributed samples is at the heart of many statistical models. The i.i.d. assumption may be reasonable when

- Data are collected from a formal sampling scheme (e.g. simple random sample) or well-designed experiment in which randomization is used;
- Data are not likely to be systematically similar because of proximity in time or space; and
- Data are not taken from clusters of observational units that might be similar.

The i.i.d. assumption

Examples:

- Temperature measurements are taken every 30-minutes for 24 hours. These samples are not i.i.d. because the measurements that are taken at nearby time points are likely to be more similar to each other than measurements taken many hours apart.
- A device that simulates a cannulation procedure is developed for training of medical personnel. Five volunteer participants complete the procedure while a sensor tracks their movements. Each volunteer performs the task ten times. The 50 measurements obtained from this experiment are (likely) not i.i.d. because the repeated measurements from one volunteer will be more similar to each other than to the measurements from a different volunteer.

Where do statistical models come from?

A researcher or data analyst specifies a statistical model based on features of the data and/or knowledge about the population being studied.

- The model will always be an approximation to the truth.
- All models are wrong; some are useful. - George Box (paraphrase)
- Most models have a close tie to existing procedures for statistical inference. This means that in order to use the procedure, you need to be willing to accept its corresponding model as a reasonable approximation to the truth.

Where do statistical models come from?

- Most models have a close tie to existing procedures for statistical inference. This means that in order to use the procedure, you need to be willing to accept its corresponding model as a reasonable approximation to the truth.

Example

A method called a two-sample t test can be used to determine if two groups have different means.

- This test is derived from the "normal, two means" model.
- In order to use this test, we need to verify that this model is reasonable as an approximate distribution of the sampled values.

For the remainder of the course, we will discuss diagnostic methods for checking whether the modeling assumptions are reasonable.