Statistical Inference: Part I

Andrew Stokes

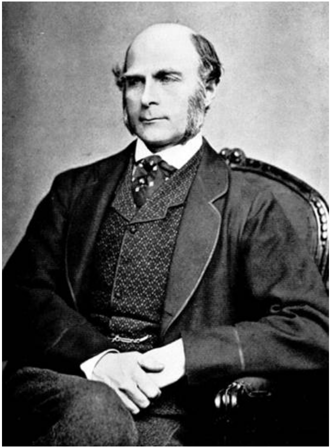
April 11, 2017

Statistics: no longer just for statisticians

Use the tools and methods of statistics is no longer restricted to statisticians. A number of factors have enabled this transformation.

- More emphasis on statistics in secondary education
- Rise in computer literacy
- Exponential increases in computing power
- User-friendly software for conducting data analyses
- Ubiquity of data
- Empirical applications vs. theory
- More emphasis on simulation and less on math

Then: Sir Francis Galton



Now: Hadley Wickham

Now: Increasing gender diversity

Local

Women flocking to statistics, the newly hot, high-tech field of data science

A | 🖶 | 11

Erin Blankeship, left, statistics professor at University of Nebraska- Lincoln, and Aimee Schwab, graduate teaching assistant and PhD student in statistics, in a classroom at Hardin Hall. Statistics is leading all other STEM fields in in attracting, retaining and promoting women. (Jake Crandall/For The Washington Post)

Today

- Confidence Intervals
- Statistical Tests
- Correlation
- Linear Regression

Type of outcome measure[1]

- Proportions (e.g. proportion of vaccinated subjects who develop a protective level of antibodies)
- Rates (e.g. incidence rate of relapse following treatment)
- Means (e.g. mean packed cell volume (PCV) at the end of malaria season)

[1] Subsequent slides based on Smith text (2015)

Sampling error

- Uncertainty arises when your data consist of a sample rather than say a census of the entire population
- This uncertainty is known as *sampling error*
- Differs from non-sampling errors (e.g. scale for weighing patients is biased)

Precision

- Statistical inference allows us to draw conclusions about the true value of the outcome measure based on sample
- The observed value of the outcome measure generally gives the best estimate of the true value
- It is also useful to have an indication of precision of this estimate
- This is done by calculating a *confidence interval*

Confidence intervals

- The CI is a range of plausible values for the true value of the outcome measure.
- Usually use a 95% CI
- Calculated so that there is a 95% probability that the CI includes the true value of the outcome measure

Confidence intervals

- Suppose true value of the outcome measure is σ and it is estimated from the sample data as $\hat{\sigma}$
- The 95% CIs will be of the form $\hat{\sigma} \pm 1.96 * SE(\hat{\sigma})$
- Here $SE(\hat{\sigma})$ denotes the *standard error* of the estimate
- Measure of the amount of sampling error
- The larger the sample size the smaller the standard error and thus the narrower the CI

The normal distribution

- The value 1.96 for calculating the 95% CI is derived from the normal distribution
- In this distribution, 95% of values are expected to fall within 1.96 SD of the mean
- For a 90% CI, the multiplying factor is 1.64

Statistical tests

- Used to test a specific hypothesis about an outcome measure
- The *null hypothesis* is often that there is true difference between the outcomes in the groups under comparison
- Did the observed difference arise just by chance, due to sampling error?

Statistical tests

- Sample data are used to calculate a *test statistic*, which gives a measure of the difference between groups
- Once calculated, its value is used to determine the p-value or statistical significance of the results
- The p-value measures the probability of obtaining a value for the statistic as extreme as the one actually observed if the null hypothesis were true
- So a very low p-value indicates that the null hypothesis is likely to be false

Example: Malaria Vaccine Trial

- Efficacy of vaccine was found to be 20%, with an associated p-value of 0.03
- This means that under the null hypothesis (vaccine had a true efficacy of zero), there would be only a 3% chance of obtaining and observed efficacy of 20% or greater

p-values

- Smaller the p-value the less plausible the null hypothesis
- A p-value of 0.001 implies that the null hypothesis is highly implausible
- In this case we have very strong evidence of a real difference between groups
- However, a p-value of 0.20 implies that a difference of the observed magnitude could have occurred by chance, even if there were no real difference between the groups

p-values

- p-values of 0.05 and below are typically considered reasonable evidence against the null hypothesis
- Results below this threshold are referred to as indicating a *statistically significant difference*
- Always preferable to report actual p-values (as opposed to stars)

p-values

- A small p-value is evidence for a real difference between groups
- BUT, a larger non-significant p-value does not indicate no difference
- Rather, it indicates that there is insufficient evidence to reject the null hypothesis
- It is never possible to prove the null hypothesis

Confidence intervals vs. statistical tests

- Statistical tests aren't everything
- Usually more important to estimate the difference and to specify a CI around it
- This provides an indication of the plausible range of differences
- This may include a zero difference

Confidence interval for a single proportion

Use analysis of proportions when the outcome is binary variable and a proportion can be calculated across individuals in the sample. The standard error of a proportion p, calculated from a sample of n subjects is estimated as

$$SE(p) = \sqrt{\left[\frac{p(1-p)}{n}\right]}$$

The 95% CI for a proportion is then given by $p \pm 1.96 * SE(p)$

Difference between two proportions

Now take a proportion that you would like to compare across two groups of individuals. The standard error of the difference between two proportions p_1 and p_2 based on n_1 and n_2 observations, is estimated as,

$$SE(p_1 - p_2) = \sqrt{\left[\bar{p}(1-\bar{p})[\frac{1}{n_1} + \frac{1}{n_2}]\right]}$$

where $\bar{p} = \frac{(n_1 p_1 + n_2 p_2)}{n_1 + n_2}$. The 95% CI for the idfference between proportions is given by $(p_1 - p_2) \pm 1.96 * SE$.

Difference between two proportions

To test the null hypothesis that there is no true difference between the two proportions, consider the following 2x2 table.

Table 21.2 Comparison of two proportions

Group	Outcome		Total	Proportion with outcome
	Yes	No		
1	a (90)	b (210)	n_1 (300)	$p_1 = a/n_1$ (0.30)
2	c (135)	d (165)	n_2 (300)	$p_2 = c/n_2$ (0.45)
Total	m_1 (225)	m_2 (375)	N (600)	

Hypothesis test for diff in two proportions

In the table, a is the number in group 1 who experiences the outcome. The expected value of a, E(a), and the variance of a, V(a), are calculated under the hypothesis of no difference between the two groups:

$$E(a) = \frac{m_1 n_1}{N}$$

$$V(a) = \frac{n_1 n_2 m_1 m_2}{N^2(N-1)}$$

Hypothesis test for diff in two proportions

The chi-squared χ^2 can then be calculated. This measures how much the observed data differ from those expected if the two proportions were truly equal.

$$\chi^2 = \frac{(|a - E(a)| - 0.5)^2}{V(a)}$$

Hypothesis test for diff in two proportions

- The χ^2 value is compared to a table of the chi-squared distribution with one degree of freedom (df).
- If it exceeds 3.84, then $p<0.05$, indicating some evidence of a real difference in the proportions.
- If any of the quantities (E(a), E(b), etc.) are less than 5.0 and N is less than 40, an alternative test should be used: 'Fisher's exact test'.

Confidence interval for a mean

- For a mean $\bar{x}$ of a sample of n observations, the standard error of the mean is given by $\bar{x} = \frac{\sigma}{\sqrt{n}}$
- Here σ is the standard deviation of the variable measured.
- The 95% CI on the mean is given by $\bar{x} \pm 1.96(\frac{\sigma}{\sqrt{n}})$

Confidence interval for a mean

- The standard deviation in the population is not known and thus must be estimated based on the sample data
- The estimate of σ is also subject to sampling error and this must be taken into account
- This is done using a multiplying factor from in the CI taken from the t-distribution, rather than using the Normal distribution.

Confidence interval for a mean

- The value of the factor will depend on the size of the sample
- The value will also depend on degrees of freedom (here, n-1)
- If the sample size is 30 or more, little error is introduced by using 1.96
- The 95% CI on the mean is given by $\bar{x} \pm t(\frac{s}{\sqrt{n}})$

Difference between two means

Let's say we want to compare two groups with means $\bar{x_1}$ and $\bar{x_2}$ with corresponding standard deviations s_1 and s_2. The standard error of the difference between the means is given by

$$s = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

Difference between two means

The 95% CI for the difference between the means is given by:

$$(\bar{x_1} - \bar{x_2}) \pm ts\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

Here t is taken from a table of the t-distribution with $(n_1 + n_2 - 2)$ df.

Test statistic

- To test the null hypothesis that there is no true difference in the means between the two groups, a t-test can be performed.
- A test statistic is calculated to assess the probability of the observed result (or a result even more extreme) if there really is no difference between the two groups.
- The difference in means divided by the standard error of the difference gives the value of a test statistic that can be looked up in tables of t-distribution with df as stated above.
- Easy to do in R!