

Statistics 305/605: Introduction to Biostatistical Methods for Health Sciences

Chapters 8-10: Review of Statistical Inference

Jinko Graham

2018-09-12

Goals for course

- ▶ The overarching goal is to get an idea of how to “think statistically”.
 - ▶ Specific techniques are less important than the general way of thinking.
 - ▶ Mechanically calculating t tests, or chi-squared tests, etc., as a human extension to the computer is not the point.
 - ▶ There are bigger ideas to pursue.
- ▶ Try to bring in the bigger ideas: quantifying uncertainty and avoiding bias to make inference from data.
 - ▶ Syllabus tilted towards quantifying uncertainty rather than avoiding bias, but will try to work in both throughout course.

Sampling Distributions (Chapter 8).

Sampling Distributions

- ▶ A statistic is a number that can be computed from data
- ▶ Its *sampling distribution* is the distribution we obtain by repeatedly drawing random samples of data from the population, and recalculating the statistic
- ▶ A *parameter* is a population quantity, such as the population mean.
- ▶ “Statistical inference” is using data from random samples to drawing conclusions about *parameters*.
- ▶ To make inference about a parameter, we need a *statistic* to estimate the parameter and the *sampling distribution* of the statistic.

Sampling Distribution of the Sample Mean

- ▶ Say that the population mean, μ , is the parameter that we are interested in.
- ▶ The sample mean, \bar{X} , is a statistic that estimates μ .
- ▶ The sampling distribution of \bar{X} is the distribution of \bar{x} values obtained by repeated sampling from the population.
- ▶ We use lower-case to denote an *observed value* and upper-case to denote the corresponding *random variable*.
 - ▶ Recall that a variable is *random* if its value is determined by chance according to a sampling distribution.

Online demo: http://onlinestatbook.com/stat_sim/sampling_dist/index.html

Properties of the Sampling Distribution of \bar{X}

1. Centre and spread: When a population has mean μ and SD σ , the sampling distribution of \bar{X} has mean μ and SD σ/\sqrt{n} , where n is the size of the sample drawn from the population (e.g. $n = 5$ in demo)
 - ▶ Since \bar{X} has mean μ it is said to be an unbiased estimator of μ .
 - ▶ Since the SD of \bar{X} decreases at rate $1/\sqrt{n}$ as the sample size n increases, \bar{X} is said to follow the “square-root law”.

2. Shape (The Central Limit Theorem, CLT): If the sample size n is large, the sampling distribution of \bar{X} is approximately normal regardless of the shape of the population distribution
- ▶ Unfortunately no universal rule for what is “large” n .
 - ▶ However, the CLT is a remarkable result. It tells us the approximate shape of the distribution of \bar{X} no matter the shape of the population distribution.
 - ▶ In the demo, try drawing 100K samples of size $N = 25$ in 3rd panel from a skewed population distribution in 1st panel.
 - ▶ Can see that the sampling distribution of \bar{X} has the same mean as the population distribution and that its SD is approximately the SD of the population distribution divided by $5 = \sqrt{25}$

Confidence Intervals (Chapter 9)

Confidence Intervals

- ▶ Estimates of a parameter without some indication of their variability or precision are not much use.
- ▶ Confidence intervals (CIs) attach a measure of precision to an estimate.
- ▶ Intervals are often of the form estimate \pm margin of error.
 - ▶ These are called “2-sided” and will be the only type of CI that we consider in this course.
- ▶ The confidence level states how often the interval covers the true parameter value
 - ▶ Note: The coverage is a statement about the *random interval*, not about the value of the true parameter, which is fixed.

Online demo: <http://wise.cgu.edu/portfolio/demo-confidence-interval-creation/>

CI for a Mean – Known SD

- ▶ If the SD σ is known, a level- C CI for μ is

$$\bar{x} \pm z^* \times \sigma / \sqrt{n},$$

where z^* is the upper $(1 - C)/2$ critical value of the standard-normal distribution.

- ▶ Note: The text calls this a 2-sided CI, and also discusses 1-sided intervals. We will restrict attention to 2-sided only.
- ▶ E.G. Since $z^* = 1.96$, the interval $\bar{X} \pm 1.96 \times \sigma / \sqrt{n}$ should cover the population mean μ about 95% of the time.
- ▶ The margin of error gets smaller as:
 - ▶ the confidence level or coverage probability C gets smaller (e.g. $C = 99$ decreases to $C = 95$),
 - ▶ the population standard deviation σ gets smaller,
 - ▶ the sample size n gets bigger

CI for a Mean – Unknown SD

- ▶ If SD σ unknown, estimate it by sample SD, s .
- ▶ A level- C CI for μ is then

$$\bar{x} \pm t^* \times s/\sqrt{n},$$

where t^* is the upper $(1 - C)/2$ critical value of the appropriate t distribution

- ▶ Use a t rather than a normal distribution to account for not knowing σ .
- ▶ Shape is similar to standard-normal distribution, but with heavier tails quantified by the degrees of freedom (df).
 - ▶ Small df, heavy tails; large df, light tails (and approaching a normal distribution).
- ▶ Heavier tails mean extreme observations are more probable; they account for the extra uncertainty of not knowing σ .
- ▶ The appropriate df is $n - 1$ when σ is estimated with a sample size of n data points.

Coverage of the CI

- ▶ When σ is known, the coverage of the CI can be derived from the following statement (details not shown):

$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$ is within z^* of 0 approximately $C \times 100\%$ of the time

- ▶ Similarly, when σ is unknown and is estimated by s , coverage of the CI can be derived from the statement:

$\frac{\bar{X} - \mu}{s/\sqrt{n}}$ is within t^* of 0 approximately $C \times 100\%$ of the time

- ▶ $\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$ and $\frac{\bar{X} - \mu}{s/\sqrt{n}}$ are called *pivotal quantities*.
- ▶ Pivotal quantities have a known distribution, which makes them useful for statistical inference. We'll come back to this later in the course.

Example

- ▶ Example from the text, page 223, summarizes data on plasma-aluminum levels, in $\mu g/l$, for $n = 10$ infants receiving antacids that contain aluminum.
 - ▶ The sample mean of the plasma-aluminum levels is $\bar{x} = 37.20 \mu g/l$, and the sample SD is $s = 7.13$
- ▶ To calculate a 95% CI we need the upper $(1 - C)/2 = (1 - 0.95)/2 = 0.025$ critical value for the t -distribution with $n - 1 = 9$ df.
 - ▶ Statistical software tells us that the critical value is $t^* = 2.262$ (see R demo).
- ▶ The 95% CI is therefore $\bar{x} \pm t^* \times s/\sqrt{n}$, or
$$(37.2 - 2.262 \times 7.13/\sqrt{10}, 37.2 + 2.262 \times 7.13/\sqrt{10}) \approx (32.1, 42.3)$$

Hypothesis Tests (Chapter 10)

Hypothesis Tests

- ▶ Hypothesis tests assess the evidence provided by the data against a null hypothesis, H_0 , in favour of an alternative hypothesis, H_a (denoted H_A in the text).
- ▶ Will illustrate with testing a null hypothesis about the value of a population mean
- ▶ Some key points to remember:
 - ▶ Hypotheses are phrased in terms of population parameters of interest (e.g. population means, **not sample means**)
 - ▶ The alternative hypothesis can be one-sided or two-sided
 - ▶ To assess the evidence for or against a hypothesis in a sample of data, we use a test statistic with a known sampling distribution under the null hypothesis
 - ▶ Extreme values of the test statistic are taken as evidence against the null hypothesis in favour of the alternative hypothesis.

Hypotheses for a Population Mean

- ▶ Consider a particular value μ_0 (e.g., $\mu_0 = 0$) for the population mean μ .
- ▶ A **two-sided** alternative hypothesis is $H_a : \mu \neq \mu_0$
- ▶ **One-sided** alternative hypotheses are $H_a : \mu > \mu_0$ or $H_a : \mu < \mu_0$.

Test Statistics

- ▶ When σ is known, we base the test on the “z statistic”

$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}.$$

- ▶ Under $H_0 : \mu = \mu_0$, Z has a standard-normal distribution, written as $Z \sim N(0, 1)$.
- ▶ When σ is unknown, we base the test on the “t statistic”

$$T = \frac{\bar{X} - \mu_0}{s/\sqrt{n}}.$$

- ▶ Under H_0 , T has a t distribution with $n - 1$ df, written as $T \sim t_{n-1}$.
- ▶ For both statistics, numerator is difference between sample mean and hypothesized value; denominator is SD of this difference.
 - ▶ Looking at the difference in the context of its variability.
 - ▶ Measures of “signal to noise”. Values of 1000 are more impressive than values of 1.

Extreme Values of the Test Statistic

- ▶ Observed values of Z or of T that are unlikely under H_0 , and more compatible with H_a than with H_0 , are taken as evidence against H_0 .
- ▶ The p -value is the chance of a test statistic that is as extreme as or more extreme than what we have observed, when the null hypothesis is true.
- ▶ How we define “extreme” depends on H_a .
- ▶ Illustrate with the t -test.

Extreme Values of T when $H_a : \mu \neq \mu_0$

- ▶ $H_a : \mu \neq \mu_0$ vs. $H_0 : \mu = \mu_0$.
- ▶ Use the statistic \bar{X} as a proxy for the parameter μ
- ▶ Observed values \bar{x} that are far from the hypothesized value μ_0 are taken as evidence in favour of H_a .
- ▶ Since the numerator of the t -statistic, T , is $\bar{X} - \mu_0$, this is equivalent to T having an observed value t that is far from zero.
 - ▶ i.e. $|t|$ much greater than zero, where $|t|$ is the absolute value of t .

p -values

- ▶ The p -value, p , is the chance that the test statistic is *as or more extreme than* the observed value given that $H_0 : \mu = \mu_0$ is true.
 - ▶ For the t -test, the statistic is T , a random variable having a t -distribution on $n - 1$ df, and the observed value is t
- ▶ From the above discussion of what “extreme” means, we can argue that, for

$$H_a : \mu \neq \mu_0,$$

the pvalue is $p = 2P(T \geq |t|)$, where

- ▶ $P(A)$ is the probability of event A ,
- ▶ t is the observed t -statistic, and
- ▶ $|t|$ is the absolute value of t .

Example

- ▶ Return to example on plasma-aluminum levels in infants.
- ▶ In the **population** of infants **not taking antacids**, the mean plasma-aluminum levels are known to be $\mu_0 = 4.13 \text{ } \mu\text{g/l}$.
- ▶ Want to assess whether the mean level μ in infants taking antacids is the same as μ_0 ; i.e. whether

$$H_0 : \mu = 4.13.$$

- ▶ Our alternative hypothesis is $H_a : \mu \neq 4.13$; i.e., the mean plasma-aluminum levels of infants taking antacids is different from infants not taking antacids.

Details

- ▶ In the sample of 10 infants taking antacids, the plasma-aluminum levels had sample mean and sample SD of $\bar{x} = 37.20 \mu\text{g/l}$ and $s = 7.13$, respectively.
- ▶ The t -statistic is

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{37.20 - 4.13}{7.13/\sqrt{10}} \approx 14.67$$

- ▶ The p -value is $2P(T \geq |14.67|)$ for T with $10 - 1 = 9$ df, which computer software reports to be $1.36833\text{e-}07$.
 - ▶ What??? This is the computer representation of 1.36833×10^{-7} , or 0.000000136833
 - ▶ Move the decimal point 7 places to the left of its current position in 1.36833.

Hypothesis Tests

- ▶ Small p -values are evidence against H_0 , in favour of H_a .
- ▶ We may set a level α in advance that marks the point at which evidence against H_0 is considered strong enough to “reject” it.
 - ▶ If we don't reject the H_0 we **retain** it.
 - ▶ **Caution:** Some people are bothered by ‘accept H_0 ’. Best say ‘retain H_0 ’ to avoid offending them (and losing marks).
- ▶ If the p -value is less than α , we “reject H_0 at level α ”.
- ▶ Historically, such a test result was described as “statistically significant”, but this terminology is going out of style.
 - ▶ In part because statistical significance says nothing about practical significance.
 - ▶ For example, in a large clinical trial with thousands of subjects, a drug may lower cholesterol ever-so-slightly, to the point of statistical significance (because the trial is so large), but without any clinical significance.

Hypothesis Testing Errors and Error Rates

- ▶ Important point: Statistical hypothesis testing can make errors.
- ▶ The “confusion matrix” for the true state of nature (rows) and the action of a hypothesis test (columns) is

	Reject H_0	Retain H_a
null (H_0)	false positive	true negative
alternative (H_a)	true positive	false negative

- ▶ False positives and negatives are called type-I and II errors, respectively.

False-Positive or Type-I Error Rate

- ▶ Mistakenly rejecting H_0 (i.e., rejecting H_0 when it is true) is a **type-I error**.
- ▶ The probability of making a type-I error, known as the **type-I error rate**, is written

$$P(\text{reject } H_0 \mid H_0 \text{ true}),$$

where $P(A \mid B)$ denotes probability of event A given event B .

(more on probability soon)

False-Negative or type-II Error Rate

- ▶ We define the false-negative error rate for completeness, but will not use it in this course.
- ▶ Mistakenly retaining H_0 (i.e., retaining H_0 when it is false) is a type-II error.
- ▶ The probability of making a type-II error, known as the **type-II error rate**, is written

$$P(\text{retain } H_0 \mid H_0 \text{ false}),$$

and is denoted by β .

Statistical Power

- ▶ As with type-II error, we define **power** but do not use it in this course.
- ▶ The **power** of a test is the probability that it correctly rejects H_0 ; i.e., the probability that the test rejects H_0 when H_0 is false:

$$P(\text{reject } H_0 \mid H_0 \text{ false}) \text{ or } 1 - \beta,$$

where β is the type-II error rate.

- ▶ For a given alternative hypothesis, one can compute the sample size required to achieve a given power (text, Section 10.6).
- ▶ Such “sample-size calculations” are frequently required by health-funding agencies in proposals for research studies.

Summary of Statistical Inference of a Population Mean

- ▶ Statistical inference: Learning about population parameters from data of a random sample from the population that are subject to random variation.
- ▶ Key point: We know the (approximate) distribution of pivotal quantities such as

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}, \sigma \text{ known; or } \frac{\bar{X} - \mu}{s/\sqrt{n}}, \sigma \text{ unknown;}$$

regardless of the shape of the population distribution.

- ▶ This result relies on the CLT, which tells us that sample averages such as \bar{X} are approximately normally distributed.
 - ▶ Many of the statistics we will study are based on averages, so inference of a population mean is a useful template.
- ▶ Knowing the distribution of the pivotal quantity allows us to construct confidence intervals, calculate p -values, test statistical hypotheses, calculate power, etc.