Elements of statistics (MATH0487-1)

Prof. Dr. Dr. K. Van Steen

University of Liège, Belgium

November 19, 2012

Outline I

1. Introduction to Statistics
   - Why?
   - What?
     - Probability
     - Statistics
   - Some Examples
   - Making Inferences
     - Inferential Statistics
     - Inductive versus Deductive Reasoning
2. Basic Probability Revisited
3. Sampling
   - Samples and Populations
   - Sampling Schemes
     - Deciding Who to Choose
     - Deciding How to Choose
     - Non-probability Sampling
     - Probability Sampling
   - A Practical Application
   - Study Designs

Outline II

  - Classification
  - Qualitative Study Designs
- Popular Statistics and Their Distributions
- Resampling Strategies

4 Exploratory Data Analysis - EDA
- Why?
  - Motivating Example
- What?
  - Data analysis procedures
  - Outliers and Influential Observations
- How?
  - One-way Methods
  - Pairwise Methods
- Assumptions of EDA

5 Estimation
- Introduction
- Motivating Example
- Approaches to Estimation: The Frequentist's Way
- Estimation by Methods of Moments

Outline III

- Motivation
- What?
- How?
- Examples
- Properties of an Estimator

- Recapitulation
  - Point Estimators and their Properties
  - Properties of an MME
- Estimation by Maximum Likelihood
  - What?
  - How?
  - Examples
  - Profile Likelihoods
  - Properties of an MLE
  - Parameter Transformations

6 Confidence Intervals

- Importance of the Normal Distribution
- Interval Estimation
  - What?
  - How?
  - Pivotal quantities

Outline IV

- Examples
- Interpretation of CIs
- Recapitulation
  - Pivotal quantities
  - Examples
  - Interpretation of CIs

7 Hypothesis Testing
- General procedure
- Hypothesis test for a single mean
- P-values
- Hypothesis tests beyond a single mean
- Errors and Power

8 Table analysis
- Dichotomous Variables
  - Comparing two proportions using2 × 2 tables
  - About RRs and ORs and their confidence intervals
- Chi-square Tests
  - Goodness-of-fit test
  - Independence test

Outline V

- Homogeneity test

Pivotal Quantity

- A **pivotal quantity or pivot** is generally defined as a function of observations and unobservable parameters whose probability distribution does not depend on unknown parameters
- Any probability statement of the form

$$P(a < H(X_1, X_2, \ldots, X_n; \theta) < b) = 1 - \alpha$$

  will give rise to a probability statement about θ
- Hence, pivots are crucial to construct confidence intervals for parameters of interest.
- Examples when sampling from a normal distribution:
  - $z = \frac{\overline{X} - \mu}{\sigma / \sqrt{n}}$ (population variance known)
  - $t = \frac{\overline{X} - \mu}{s / \sqrt{n}}$ (population variance unknown)

Confidence intervals for means

Towards a summary table:

Normal Pop. & Population Var. known

$$P(a < H(X_1, X_2, \ldots, X_n; \theta) = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} < b) = 1 - \alpha$$

Normal Pop. & Population Var. UNknown

$$P(a < H(X_1, X_2, \ldots, X_n; \theta) = \frac{\overline{X} - \mu}{s/\sqrt{n}} < b) = 1 - \alpha$$

Binom. Pop. & Sample Size Large

$$P(a < H(X_1, X_2, \ldots, X_n; \theta) = \frac{\hat{p} - P}{\sqrt{\hat{p}(1-\hat{p})}/\sqrt{n}} < b) = 1-\alpha$$

Confidence intervals for means

Summary table:

Population Distribution	Sample Size	Population Variance	95% Confidence Interval
Normal	Any Any	σ^2 known σ^2 unknown, use s^2	$\bar{X} \pm 1.96\sigma/\sqrt{n}$ $\bar{X} \pm t_{0.025,n-1}s/\sqrt{n}$
Not Normal/ Unknown	Large Large Small	σ^2 known σ^2 unknown, use s^2 Any	$\bar{X} \pm 1.96\sigma/\sqrt{n}$ $\bar{X} \pm 1.96s/\sqrt{n}$ Non-parametric methods
Binomial	Large Small	- -	$\hat{p} \pm 1.96\sqrt{\hat{p}(1-\hat{p})/n}$ Exact methods

Confidence intervals for difference in means

- Recall that formula's for CIs for a single mean depend on
  - whether or not σ^2 is known
  - sample size
- For a difference in means, the formula's for CIs depend on
  - whether or not the variances are assumed to be equal when the variances are unknown
  - sample size in each group

Confidence intervals for difference in means

Variances assumed to be equal:

- The standard error of the difference is estimated by

$$\sqrt{\frac{s_p^2}{n_1}+\frac{s_p^2}{n_2}},$$

- with s_p^2 the pooled variance

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

n_1 and n_2 the sample sizes of sample 1 and 2 respectively, or rewritten:

$$s_p^2=\frac{\nu_1 s_1^2+\nu_2 s_2^2}{\nu_1+\nu_2},$$

where $\nu_1=n_1-1$ and $\nu_2=n_2-1$.

Is Unbiasedness a Good Thing?

- Unbiasedness is important when combining estimates, as averages of unbiased estimators are unbiased.
- Example:
  - When combining standard deviations $s_1, s_2, \ldots, s_k$ with degrees of freedom $\mathrm{df}_1, \ldots, \mathrm{df}_k$ we always average their squares

$$\bar{s} = \sqrt{\frac{\mathrm{df}_1 s_1^2 + \cdots + \mathrm{df}_k s_k^2}{\mathrm{df}_1 + \cdots + \mathrm{df}_k}}$$

    as s_i^2 are unbiased estimators of the variance σ^2, whereas s_i are not unbiased estimators of σ.

Therefore, be careful when averaging biased estimators! It may well be appropriate to make a bias-correction before averaging.

Confidence intervals for difference in means

Variances assumed to be equal:

- In this case, it is clear that the standard error can be estimated more efficiently by combining the samples
- Hint: Assume that the new estimator s^2 is a linear combination of the sample variances s_1^2 and s_2^2 such that s^2 has the smallest variance of all such linear, unbiased estimators.
- Then, when we write $s^2 = a_1 s_1^2 + a_2 s_2^2$, it can be shown that

$$Var(s^2) = a_1^2 Var(s_1^2) + (1 - a_1)^2 Var(s_2^2),$$

and

$$a_1 = \frac{Var(s_2^2)}{Var(s_1^2) + Var(s_2^2)},$$

$$a_2 = \frac{Var(s_1^2)}{Var(s_1^2) + Var(s_2^2)}$$

Confidence intervals for difference in means

Variances assumed to be UNequal:

- The standard error of the difference is estimated by

$$\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}.$$

- When sampling from two normal populations $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$, with $\sigma_1 \neq \sigma_2$ and unknown, then

$$t = \frac{(\overline{X_1} - \overline{X_2}) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Any clue about the degrees of freedom?

Confidence intervals for difference in means

Variances assumed to be UNequal:

- The degrees of freedom $df = \nu$ is taken to be

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

- This formula was developed by the statistician Franklin E. Satterthwaite. The motivation and the derivation of the df result is given in Satterthwaite's article in Psychometrika (vol. 6, no. 5, October 1941), for those interested - no exam material

Confidence intervals for difference in means

Variances assumed to be UNequal:

- It is “safe” to adopt equal variance procedures when $\frac{s_2}{s_1} < 2$ (s_2 the larger one)
- The problem of unequal (unknown) variances is known as the Behrens-Fisher problem and various solutions have been given (beyond the scope of this course)
- When unsure to consider the unequal variance solution, take it. It may be the most conservative choice (less “power” - see later Chapter “Hypothesis Testing”), but it will be the choice less likely to be incorrect.

Confidence intervals for difference in means

Summary table:

Population Distribution	Sample Size	Population Variances	95% Confidence Interval
Normal	Any	known	$(\bar{X}_1 - \bar{X}_2) \pm 1.96\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$
	Any	unknown, $\sigma_1^2 = \sigma_2^2$	$(\bar{X}_1 - \bar{X}_2) \pm t_{0.025, n_1+n_2-2}\sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}$
	Any	unknown, $\sigma_1^2 \neq \sigma_2^2$	$(\bar{X}_1 - \bar{X}_2) \pm t_{0.025, \nu}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$
Not Normal/ Unknown	Large	known	$(\bar{X}_1 - \bar{X}_2) \pm 1.96\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$
	Large	unknown, $\sigma_1^2 = \sigma_2^2$	$(\bar{X}_1 - \bar{X}_2) \pm 1.96\sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}$
	Large	unknown, $\sigma_1^2 \neq \sigma_2^2$	$(\bar{X}_1 - \bar{X}_2) \pm 1.96\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$
	Small	Any	Non-parametric methods

Confidence intervals for difference of proportions

Summary table:

Population Distribution	Sample Size	95% Confidence Interval
Binomial	Large	$(\hat{p}_1 - \hat{p}_2) \pm 1.96\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$
	Small	Exact methods

Pivotal Quantities with Chi-square distribution

Theorem

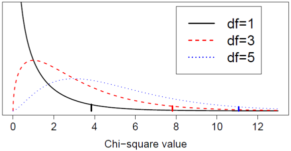
If $Z_1, \ldots, Z_n$ is a random sample from a standard normal distribution, then

1. $\overline{Z}$ *has a normal distribution with mean* 0 *and variance* $1/n$
2. $\overline{Z}$ *and* $\sum_{i=1}^{n}(Z_i - \overline{Z})^2$ *are independent*
3. $\sum_{i=1}^{n}(Z_i - \overline{Z})^2$ *has a chi-square distribution with* $n-1$ *degrees of freedom*

- The aforementioned theorem gives results to remember!
- The special case of $Z_i = \frac{X_i - \mu}{\sigma}$, gives $\overline{Z} = \frac{\overline{X} - \mu}{\sigma}$, and $\sum_{i=1}^{n}(Z_i - \overline{Z})^2 = \sum_{i=1}^{n}((X_i - \overline{X})^2/\sigma^2)$

Confidence intervals for variances

- Similarly, any random variable $U = \sum_{i=1}^{n} (X_i - \mu)^2/\sigma^2$ with $X_1, \ldots, X_n$ representing a random sample from a normal distribution with mean μ and variance σ^2, has a **chi-square** distribution with n degrees of freedom.
- The formula's for CIs for a single variance depend on:
  - whether or not the population mean μ is known
  - sample size
- Rather than relying on a normal distribution, the chi-square distribution is a better choice here



Confidence intervals for variances

Towards a summary table:

Normal Pop. & Population Mean known

$$P(a < H(X_1, X_2, \ldots, X_n; \theta) = \sum_{i=1}^{n}(X_i - \mu)^2/\sigma^2 < b) = 1 - \alpha$$

Normal Pop. & Population Mean UNknown

$$P(a < H(X_1, X_2, \ldots, X_n; \theta) = \sum_{i=1}^{n}(X_i - \overline{X})^2/\sigma^2 < b) = 1 - \alpha$$

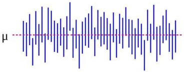
Confidence intervals for variances

Summary table:

Population Distribution	Sample Size	Population Mean	Pivot and Distribution
Normal	Any	μ known unbiased estimator of σ^2 is $s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2$	$\frac{ns^2}{\sigma^2} \sim \chi^2_n$
Normal	Any	μ unknown unbiased estimator of σ^2 is $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$	$\frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{n-1}$
Not Normal/ Unknown	Large	μ known unbiased estimator of σ^2 is $s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2$	$\frac{ns^2}{\sigma^2} \sim \chi^2_n$
Not Normal/ Unknown	Large	μ unknown unbiased estimator of σ^2 is $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$	$\frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{n-1}$
Not Normal/ Unknown	Small	Any	Non-parametric methods

Interpretation of Confidence Interval (CI)

- *Before* the data are observed, the probability is at least $(1-\alpha)$ that $[L, U]$ will contain the population parameter [Note that here, L and U are random variables]
- In *repeated sampling* from the relevant distribution, $100(1-\alpha)\%$ of all intervals of the form $[L, U]$ will include the true population parameter



- *After* the data are observed, the constructed interval $[L, U]$ either contains the true parameter value or it does not (there is no longer a probability involved here!)

A statement such as $P(3.5 < \mu < 4.9) = 0.95$ is **incorrect** and should be replaced by **A 95% confidence interval for μ is (3.5,4.9)**

Testing Hypotheses

- **Inferential statistics**

  At best, we can only be confident in our statistical assertions, but never certain of their accuracy.

- **Trying to Understand the True State of Affairs**

  In the absence of prior knowledge about the details of some population of interest, sample data serve as our best estimate of that population.

- **True State of Affairs + Chance = Sample Data**

  The laws of chance combined with the true state of affairs create a natural force that is always operating on the sampling process. Consequently, the means of different samples taken from the same population are expected to vary around the "true" mean just by chance.

Testing Hypotheses

- **Sampling Distributions**

  Populations, which are distributions of individual elements, give rise to sampling distributions, which describe how collections of elements are distributed in the population.

- **The Standard Error: A Measure of Sampling Error**

  We have some control over sampling error because sample size determines the standard error (variability) in a sampling distribution.

- **Theoretical Sampling Distributions as Statistical Models of the True State of Affairs**

  Theoretical sampling distributions have been generated so that researchers can estimate the probability of obtaining various sample means from a pre-specified population (real or hypothetical).

Towards a formalism

Making Formal Inferences about Populations: Hypothesis testing

- When there are many elements in the sampling distribution, it is always possible to obtain a rare sample (e.g., one whose mean is very different from the true population mean).
- The probability of such an outcome occurring just by chance is determined by the particular sampling distribution specified in the null hypothesis.
- When the probability (p) of the observed sample mean occurring by chance is really low (typically less than one in 20, e.g., $p < 0.05$), the researcher has an important decision to make regarding the hypothesized true state of affairs. One of two inferences can be made:
  - #1: The hypothesized value of the population mean is correct and a rare outcome has occurred just by chance.
  - #2: The true population mean is probably some other value that is more consistent with the observed data. Reject the null hypothesis in favor of some alternative hypothesis.

Towards a formalism

- The rational decision is to assume #2, because the observed data (which represent direct (partial) evidence of the true state of affairs), are just too unlikely if the hypothesized population is true.
- Thus, rather than accept the possibility that a rare event has taken place, the statistician chooses the more likely possibility that the hypothesized sampling distribution is wrong.
- However, rare samples do occur, which is why statistical inference is always subject to error.
- Even when observed data are consistent with a hypothesized population, they are also consistent with many other hypothesized populations. It is for this reason that the hypothesized value of a population parameter can never be proved nor disproved from sample data.

Towards a formalism

The nature of making inferences based on random sampling:

- We use inferential statistics to make tentative assertions about population parameters that are most consistent with the observed data. Actually, inferential statistics only helps us to rule out values; it doesn't tell us what the population parameters are. We have to infer the values, based on what they are likely not to be.
- We can make errors while doing so. Only in the natural sciences does evidence contrary to a hypothesis lead to rejection of that hypothesis without error. In statistical reasoning there is also rejection (inference #2), but with the possibility that a rare sample has occurred simply by chance (sampling error).

Basic Steps of Hypothesis Testing

- Define the null hypothesis, H_0
- Define the alternative hypothesis, H_a, where H_a is usually of the form "not" H_0, but not necessarily
- Define the type I error (probability of falsely rejecting the null), α, usually 0.05, but not necessarily
- Calculate the test statistic
- Calculate the p-value (probability of getting a result "as or more extreme" than observed if the null is true)
- If the p-value is $\leq \alpha$, **reject** H_0. Otherwise, **fail to reject** H_0

Hypothesis test for a single mean

Birthweight example

- Assume a population of normally distributed birth weights with a known standard deviation, $\sigma = 1000$ grams
- Birth weights are obtained on a sample of 10 infants; the sample mean is calculated as 2500 grams
- Question: Is the mean birth weight in this population different from 3000 grams?
- Set up a two-sided test of

$$H_0 : \mu = 3000,$$

$$H_a : \mu \neq 3000$$

- Let the probability of falsely rejecting the null be $\alpha = 0.05$

Hypothesis test for a single mean

Birthweight example

- Calculate the test statistic:

$$Z_{obs} = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} = \frac{2500 - 3000}{1000/\sqrt{10}} = -1.58$$

Do you recognize this statistic? Can you give another name for it? Can you give an interpretation for the test value?

Hypothesis test for a single mean

Birthweight example

- Calculate the test statistic:

$$z_{obs} = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} = \frac{2500 - 3000}{1000/\sqrt{10}} = -1.58$$

- Meaning: The observed mean is 1.58 standard errors below the hypothesized mean
- The test statistic is the standardized value of our data, *assuming that the null hypothesis is true*
- The question now is: If the true mean is 3000 grams, is our observed sample mean of 2500 “common” or is this value (highly) unlikely to occur?

Hypothesis test for a single mean

Birthweight example

- Calculate the p-value to answer our question:

$$p - value = P(Z \leq -|z_{obs}|) + P(Z \geq |z_{obs}|) = 2 \times 0.057 = 0.114$$

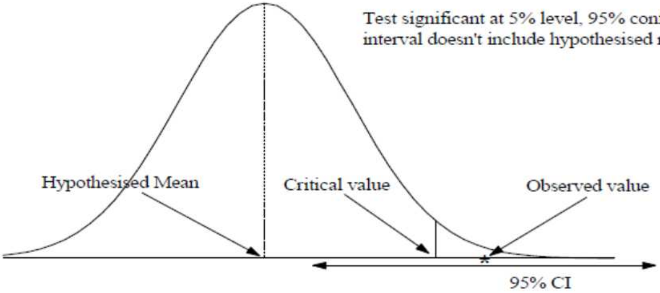
- If the true mean is 3000 grams, our data or data more extreme than ours would occur in 11 out of 100 studies (of the same size, n=10)
- In other words, in 11 out of 100 studies with sample size n = 10, just by chance we are likely to observe a sample mean of 2500 or more extreme if the true mean is 3000 grams
- What does this say about our hypothesis? We fail to reject the null hypothesis since we chose $\alpha = 0.05$ and our p-value is 0.114
- General guideline: if p-value $\leq \alpha$, then reject H_0

Other approaches for 2-sided hypothesis testing

p-value : Calculate the test statistic (TS), get a p-value from the TS and then reject the null hypothesis if p-value $\leq \alpha$ or fail to reject the null if p-value $> \alpha$

Critical Region : Alternate, equivalent approach: calculate a critical value (CV) for the specified α, compute the TS and reject the null if $|TS| > |CV|$ saying that the p-value is $< \alpha$ and fail to reject the null if $|TS| < |CV|$ saying p-value $> \alpha$. You never calculate the actual p-value.

Other approaches for 2-sided hypothesis testing



Test significant at 5% level, 95% confidence interval doesn't include hypothesised mean.
Hypothesised Mean
Critical value
Observed value
95% CI

Other approaches for 2-sided hypothesis testing

Birthweight example

- You can also use the **critical value** approach
- Based on our significance level ($\alpha = 0.05$) and assuming H_0 is true, how "far" does our sample mean have to be from $H_0 : \mu = 3000$ in order to reject?
- Critical value $= z_c$ where $2 \times P(Z > |z_c|) = 0.05$
- In our example, $z_c = 1.96$ and test statistic $z_{obs} = -1.58$
- The **rejection region** is any value of our test statistic that is ≤ -1.96 or ≥ 1.96
- $|z_{obs}| < |z_c|$ since $|-1.58| < |1.96|$, so we fail to reject the null with p-value > 0.05
- Decision is the same whether using the p-value or critical value

Other approaches for 2-sided hypothesis testing

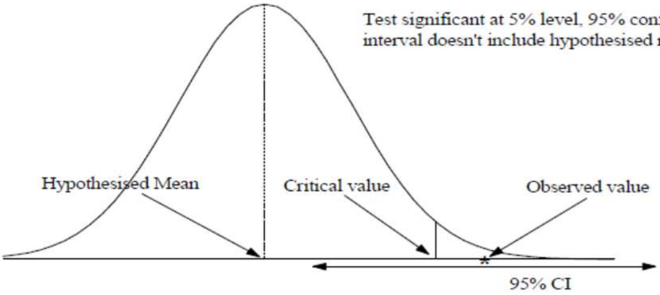
Confidence interval : Another equivalent approach goes as follows: create a $100(1-\alpha)\%$ CI for the population parameter.

- If the CI contains the null hypothesis, you fail to reject the null hypothesis with p-value $> \alpha$.
- If the CI does not contain the null hypothesis, you reject the null hypothesis with p-value $\leq \alpha$.

You never calculate the actual p-value.

The confidence interval approach does not work with one-sided test but the critical value and p-value approaches do

Other approaches for 2-sided hypothesis testing


Test significant at 5% level, 95% confidence
interval doesn't include hypothesised mean.
Hypothesised Mean
Critical value
Observed value
95% CI

Confidence intervals for means

Summary table:

Population	Sample	Population	95% Confidence
Distribution	Size	Variance	Interval
Normal	Any	σ^2 known	$\bar{X} \pm 1.96\sigma/\sqrt{n}$
	Any	σ^2 unknown, use s^2	$\bar{X} \pm t_{0.025,n-1}s/\sqrt{n}$
Not Normal/ Unknown	Large	σ^2 known	$\bar{X} \pm 1.96\sigma/\sqrt{n}$
	Large	σ^2 unknown, use s^2	$\bar{X} \pm 1.96s/\sqrt{n}$
	Small	Any	Non-parametric methods
Binomial	Large	-	$\hat{p} \pm 1.96\sqrt{\hat{p}(1-\hat{p})/n}$
	Small	-	Exact methods

Hypotheis test for a single mean

Birthweight example

- An alternative approach for two sided hypothesis testing is to calculate a $100(1-\alpha)\%$ **confidence interval for the mean** μ
- $$\hat{X} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{10}} \rightarrow 2500 \pm 1.96\frac{1000}{\sqrt{10}}$$
- The hypothetical true mean 3000 is a plausible value of the true mean given our data since it is in the CI
- We cannot say that the true mean is different from 3000
- We fail to reject the null hypothesis with p-value > 0.05
- Same conclusion as with p-value and critical value approach!

Definition of a p-value

- The **p-value for a hypothesis test** is the probability of obtaining a value of the test statistic as or more extreme than the observed test statistic when the null hypothesis is true
- The rejection region is determined by α, the desired **level of significance**, or probability of committing a type I error or the probability of falsely rejecting the null
- Reporting the p-value associated with a test gives an indication of how common or rare the computed value of the test statistic is, given that H_0 is true
- We often use z_{obs} to denote the computed value of the test statistic, since quite often we can assume a normal distribution for the test statistic of our choice

Fallacies of statistical testing

- Failure to reject the null hypothesis leads to its acceptance. (**WRONG!** Failure ro reject the null hypothesis implies insufficient evidence for its rejection.)
- The p value is the probability that the null hypothesis is incorrect. (**WRONG!** The p value is the probability of the current data or data that is more extreme assuming H_0 is true.)
- $\alpha = 0.05$ is a standard with an objective basis. (**WRONG!** $\alpha = 0.05$ is merely a convention that has taken on unwise mechanical use.)
- Small p values indicate large effects. (**WRONG!** p values tell you next to nothing about the size of a difference.)
- Data show a theory to be true or false. (**WRONG!** Data can at best serve to show that a theory or claim is highly unlikely.)
- Statistical significance implies importance. (**WRONG!WRONG!WRONG!** Statistical significance says very little about the importance of a relation.)

Hypothesis test for a single mean

Choosing the correct test statistic:

- Depends on population sd (σ) assumption and sample size
- When σ is known, we have a standard normal test statistic
- When σ is unknown and
  - our sample size is relatively small, the test statistic has a t-distribution and
  - our sample size is large, we have a standard normal test statistic (CLT)
- The only difference in the procedure is the calculation of the p-value or rejection region uses a t- instead of normal distribution

Hypothesis test for a single mean

$H_0 : \mu = \mu_0; H_a : \mu \neq \mu_0$

Population Distribution	Sample Size	Population Variance	Test Statistic
Normal	Any	σ^2 known	$z_{obs} = \frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}}$
	Any	σ^2 unknown uses s^2, df=n-1	$t_{obs} = \frac{\bar{X}-\mu_0}{s/\sqrt{n}}$
Not Normal/ Unknown	Large	σ^2 known	$z_{obs} = \frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}}$
	Large	σ^2 unknown uses s^2	$z_{obs} = \frac{\bar{X}-\mu_0}{s/\sqrt{n}}$
	Small	Any	Non-parametric methods

Hypothesis test for a single proportion

$H_0 : p = p_0; H_a : p \neq p_0$

Population Distribution	Sample Size	Test Statistic
Binomial	Large	$z_{obs} = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$
	Small	Exact methods

Hypothesis test for diff of 2 means

Choosing the correct test statistic:

- So far, we've been looking at only a single mean. What happens when we want to compare the means in two groups?
- We can compare two means by looking at the difference in the means
  - Consider the question: is $\mu_1 = \mu_2$?
  - This is equivalent to the question: is $\mu_1 - \mu_2 = 0$?
- The work done for testing hypotheses about single means extends to comparing two means
- Think about the pivotal quantities to construct confidence intervals: Assumptions about the two population standard deviations determine the formula to use

Hypothesis test for diff of 2 means

$H_0 : \mu_1 - \mu_2 = \mu_0; H_a : \mu_1 - \mu_2 \neq \mu_0$

Population Distribution	Sample Size	Population Variances	Test Statistic
Normal	Any	Known	$z_{obs} = \frac{(\bar{X}_1 - \bar{X}_2) - \mu_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$
	Any	unknown assume $\sigma_1^2 = \sigma_2^2$, $df = n_1 + n_2 - 2$ $s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}$	$t_{obs} = \frac{(\bar{X}_1 - \bar{X}_2) - \mu_0}{\sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}}$
	Any	unknown assume $\sigma_1^2 \neq \sigma_2^2$, $df = \nu = \frac{(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2})^2}{\frac{(s_1^2/n_1)^2}{n_1-1} + \frac{(s_2^2/n_2)^2}{n_2-1}}$	$t_{obs} = \frac{(\bar{X}_1 - \bar{X}_2) - \mu_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$

The EPREDA trial

- The EPREDA Trial: randomized, placebo-controlled trial to determine whether dipyridamole improves the efficacy of aspirin in preventing fetal growth retardation
- Pregnant women randomized to placebo (n=73) or to treatment (n=156)
- Mean birth weight was statistically significantly different in the two groups, with the mean weight in the treatment group being higher than the mean birthweight in the placebo group
  - Treatment group: 2751 (SD 670) grams
  - Placebo group: 2526 (SD 848) grams
- We now have the knowledge to reproduce this result

The EPREDA trial

Test the hypothesis:

$$\begin{aligned} H_0 : \mu_{placebo} &= \mu_{treated} \\ \text{vs. } H_a : \mu_{placebo} &\neq \mu_{treated} \end{aligned}$$

at the 5% significance level ($\alpha = 0.05$)

The data are:

Treatment	n	mean	SD
Placebo	73	2526	848
Treated	156	2751	670

The EPREDA trial

- Calculate the test statistic assuming the variances are unequal:

$$t_{obs} = \frac{(\bar{X}_p - \bar{X}_t) - \mu_0}{\sqrt{\frac{s_p^2}{n_p} + \frac{s_t^2}{n_t}}} = \frac{2526 - 2751}{\sqrt{\frac{848^2}{73} + \frac{676^2}{156}}} = -1.99$$

- The observed difference in mean birth weight comparing the placebo to treated groups is approximately 2 standard errors below the hypothesized difference of 0
- The degrees of freedom are:

$$\nu = \frac{\left(\frac{848^2}{73} + \frac{670^2}{156}\right)^2}{\frac{(848^2/73)^2}{73-1} + \frac{(670^2/156)^2}{156-1}} \approx 116$$

- Our sample size is pretty large, so the test statistic will behave similar to a standard normal variable

The EPREDA trial

- What is the p-value in this example?
  - p-value$=0.047$ using standard normal `2 *pnorm(-1.99)`
  - p-value$=0.049$ using t_{116} `2*pt(-1.99,df=116)`
- What is your decision in this case?
  - Not straightforward since p-value is very close to $\alpha = 0.05$
  - There may be a difference in birth weight comparing the two groups, there may not
  - Need to consider the practical implications
    - Is the treatment expensive?
    - Does the treatment produce adverse side effects?
    - Is the observed difference in mean birthweights **scientifically** important?
- One possible conclusion
  - 'marginally statistically significant' difference in mean birthweights
  - need to perform more studies

The EPREDA trial

- Can also give 95% confidence interval for the difference in the two means: (-446.13, -3.87)
- The CI is a plausible range of values for the true difference in birth weights comparing the placebo to treated groups
- What is your null hypothesis? No difference!
- Given this confidence interval, is “no difference (0)” a plausible value? Almost?

Type of errors in hypothesis testing: α and β

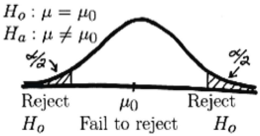
$$\begin{aligned} \beta &= P(\text{Type II error}) \\ &= P(\text{fail to reject } H_0 \text{ given } H_0 \text{ is false}) \end{aligned}$$

$$\begin{aligned} \text{Power} &= 1 - \beta \\ &= \text{probability of rejecting } H_0 \text{ when } H_0 \text{ is false} \end{aligned}$$

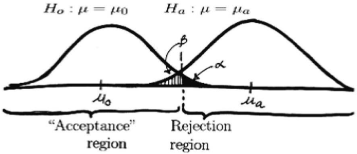
- **Aim**: to keep Type II error small and achieve large power
- β depends on sample size, α, and the specified alternative value

Type of errors in hypothesis testing: β

Two-sided hypothesis test



Power = 1 - β

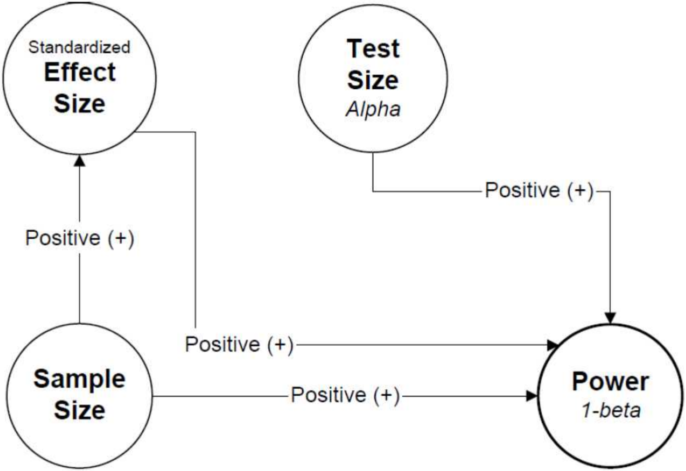


Type II error is calculated under a specified $H_a : \mu = \mu_a$

Type of errors in hypothesis testing: β

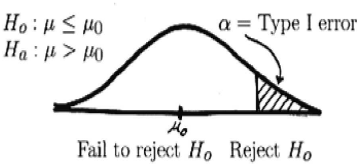
- The value of β is usually unknown since the true mean (or other parameter) is generally unknown
- Before data collection, scientists should decide on
  - the test they will perform
  - the desired Type I error rate α
  - the desired β, for a specified alternative value
- Only then can an appropriate sample size can be determined

Components of a statistical test: Relationships

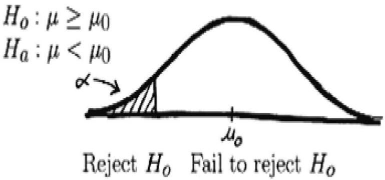


Critical Regions

One-sided hypothesis test



One-sided hypothesis test



A note aside: Effect size

- The **effect size** encodes the selected research findings on a numeric scale
- There are many different types of effect size measures (OR, difference in means, correlations,) , each suited to different research situations
- Each effect size type may also have multiple methods of computation
- An example of a **standardized effect size** ES is

$$\overline{ES} = \frac{\overline{X}_{G1} - \overline{X}_{G2}}{s_p}, s_p = \sqrt{\frac{s_1^2(n_1 - 1) + s_2^2(n_2 - 1))}{n_1 + n_2 - 2}}.$$

Does this seem natural to you?

Recall: The Standard Normal Distribution

- Definition: a normal distribution $N(\mu, \sigma^2)$ with parameters $\mu = 0$ and $\sigma = 1$
- Its density function is written as

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}, -\infty < x < \infty$$

- We typically use the letter Z to denote a **standard normal** random variable: $Z \sim N(0, 1)$
- Important: We can use the standard normal all the time (instead of non-standardized version) because if $X \sim N(\mu, \sigma^2)$ then $\frac{X-\mu}{\sigma} \sim N(0, 1)$
- This process is called "standardizing" a normal random variable

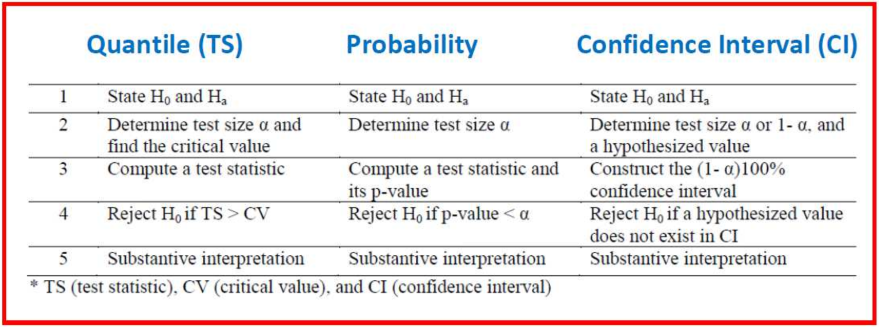
Recall: Hypothesis test for diff of 2 means

$H_0 : \mu_1 - \mu_2 = \mu_0; H_a : \mu_1 - \mu_2 \neq \mu_0$

Population Distribution	Sample Size	Population Variances	Test Statistic
Normal	Any	Known	$z_{obs} = \frac{(\bar{X}_1 - \bar{X}_2) - \mu_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$
	Any	unknown assume $\sigma_1^2 = \sigma_2^2$, $df = n_1 + n_2 - 2$ $s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}$	$t_{obs} = \frac{(\bar{X}_1 - \bar{X}_2) - \mu_0}{\sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}}$
	Any	unknown assume $\sigma_1^2 \neq \sigma_2^2$, $df = \nu = \frac{(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2})^2}{\frac{(s_1^2/n_1)^2}{n_1-1} + \frac{(s_2^2/n_2)^2}{n_2-1}}$	$t_{obs} = \frac{(\bar{X}_1 - \bar{X}_2) - \mu_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$

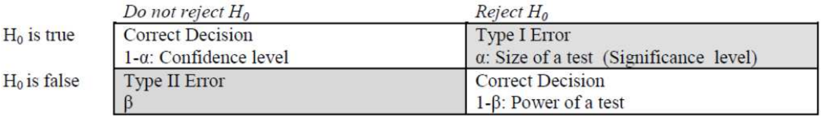
Hypothesis Testing

Bacis Summary



	Quantile (TS)	Probability	Confidence Interval (CI)
1	State H_0 and H_a	State H_0 and H_a	State H_0 and H_a
2	Determine test size α and find the critical value	Determine test size α	Determine test size α or 1- α, and a hypothesized value
3	Compute a test statistic	Compute a test statistic and its p-value	Construct the (1- α)100% confidence interval
4	Reject H_0 if TS > CV	Reject H_0 if p-value < α	Reject H_0 if a hypothesized value does not exist in CI
5	Substantive interpretation	Substantive interpretation	Substantive interpretation

* TS (test statistic), CV (critical value), and CI (confidence interval)



	Do not reject H_0	*Reject H_0*
H_0 is true	Correct Decision 1-α: Confidence level	Type I Error α: Size of a test (Significance level)
H_0 is false	Type II Error β	Correct Decision 1-β: Power of a test

Proportions and 2 × 2 tables

Population	Success	Failure	Total
Population 1	x_1	$n_1 - x_1$	n_1
Population 2	x_2	$n_2 - x_2$	n_2
Total	$x_1 + x_2$	$n - (x_1 + x_2)$	n

- Row 1 shows results of a binomial experiment with n_1 trials
- Row 2 shows results of a binomial experiment with n_2 trials

Comparing proportions

- Often, we want to compare p_1, the probability of success in population 1, to p_2, the probability of success in population 2
  - Usually: "Success" = Disease
  - Population 1 = Treatment 1
  - Population 2 = Treatment 2 (maybe placebo)
- How do we compare these proportions?
  - We've talked about comparing proportions by looking at their difference
  - But sometimes we want to look at one proportion 'relative' to the other
  - This approach depends on the type of study the data came from

Cohort Study Design

- Find a group of individuals without the disease and separate into those
  - with the exposure and
  - without the exposure

  Follow over time and measure the disease rates in both groups
- Compare the disease rate in the exposed and unexposed
- If the exposure is harmful and associated with the disease, we would expect to see higher rates of disease in the exposed group versus the unexposed group
- Allows us to estimate the incidence of the disease (rate at which new disease cases occur)

Case-control Study Design

- Identify individuals
  - with the disease of interest (case) and
  - those without the disease (control)

  and then look retrospectively (at prior records) to find what the exposure levels were in these two groups
- The goal is to compare the exposure levels in the case and control groups
- If the exposure is harmful and associated with the disease, we would expect to see higher levels of exposure in the cases than in the controls
- Very useful when the disease is rare

Study Example

Aceh Vitamin A Trial

- Exposure levels (Vitamin A) assigned at baseline and then the children are followed to determine survival in the two groups.
  - 25,939 pre-school children in 450 Indonesian villages in northern Sumatra
  - Vitamin A given 1-3 months after the baseline census, and again at 6-8 months
  - Consider 23,682 out of 25,939 who were visited on a pre-designed schedule
- References:
  1 Sommer A, Djunaedi E, Loeden A et al, Lancet 1986.
  2 Sommer A, Zeger S, Statistics in Medicine 1991.

What type of (epidemiological) design is this?

Trial outcome: 12 month mortality

	Alive at 12 months?		
Vit A	No	Yes	Total
Yes	46	12,048	12,094
No	74	11,514	11,588
Total	120	23,562	23,682

- Does Vitamin A reduce mortality?
- Calculate risk ratio or “relative risk”
  - Relative Risk abbreviated as RR
  - Could also compare difference in proportions: called “attributable risk”

Trial outcome: 12 month mortality

- In fact, with O = outcome of interest (e.g., death rate) and E = exposure (e.g., 1 if Vit A given, 2 if no Vit A given), the attributable risk in the exposed ($E = 1$) is given by

$$
\begin{aligned}
AR_E &= \frac{P(O|E=1) - P(O|E=2)}{P(O|E=1)} \\
&= \frac{(P(O|E=1) - P(O|E=2))/P(O|E=2)}{P(O|E=1)/P(O|E=2)} \\
&= \frac{RR-1}{RR}
\end{aligned}
$$

- $RR = P(O|E=1)/P(O|E=2) = p_1/p_2$

Calculation of Relative Risk (RR)

$$\begin{aligned}
\text{Relative Risk} &= \frac{\text{Rate with Vitamin A}}{\text{Rate without Vitamin A}} \\
&= \frac{\hat{p}_1}{\hat{p}_2} \\
&= \frac{46/12,094}{74/11,588} \\
&= \frac{0.0038}{0.0064} \\
&= 0.59
\end{aligned}$$

- The death rate with vitamin A is 0.60 times (or 60% of) the death rate without vitamin A.
- Equivalent interpretation: Vitamin A group had 40% lower mortality than without vitamin A group!

Confidence Interval for Relative Risk (RR)

How would you compute the variance of the RR?

Confidence Interval for Relative Risk (RR)

How would you compute the variance of the RR?

- Take natural logarithm log(RR)
- $E = 1$ and $E = 0$ groups are independent and therefore the $Var(log(RR)) = Var(log(\hat{p}_1)) + Var(log(\hat{p}_2))$
- $Var(log(\hat{p}_1)) = (1/\hat{p}_1)^2 Var(\hat{p}_1)$
- $Var(\hat{p}_1) = [\hat{p}_1(1 - \hat{p}_1)]/n_1$

Confidence Interval for Relative Risk (RR)

- Step 1: Find the estimate of the log RR

$$\log(\frac{\hat{p}_1}{\hat{p}_2})$$

- Step 2: Estimate the variance of the log(RR) as:

$$\frac{1-p_1}{n_1 p_1} + \frac{1-p_2}{n_2 p_2}$$

- Step 3: Find the 95% CI for log(RR):

$$\log(RR) \pm 1.96 \cdot SD(\log RR) = (\text{lower, upper})$$

- Step 4: Exponentiate to get 95% CI for RR;

$$e^{(\text{lower, upper})}$$

Confidence Interval for Relative Risk (RR)

95% CI for log relative risk is:

$$\begin{aligned} & \log(RR) \pm 1.96 \cdot SD(\log RR) \\ = \; & \log(0.59) \pm 1.96 \cdot \sqrt{\frac{0.9962}{46} + \frac{0.9936}{74}} \\ = \; & -0.53 \pm 0.37 \\ = \; & (-0.90, -0.16) \end{aligned}$$

95% CI for relative risk

$$(e^{-0.90}, e^{-0.16}) = (0.41, 0.85)$$

Does this confidence interval contain 1?

Suppose the data were from a case-control study

- Recall: in case-control studies, individuals are selected by outcome status
- Disease (mortality) status defines the population, and exposure status defines the success
- p_1 and p_2 have a difference interpretation in a case-control study than in a cohort study
- Cohort:
  - $p_1 = \text{P}(\text{Disease} \mid \text{Exposure})$
  - $p_2 = \text{P}(\text{Disease} \mid \text{No Exposure})$
- Case-Control:
  - $p_1 = \text{P}(\text{Exposure} \mid \text{Disease})$
  - $p_2 = \text{P}(\text{Exposure} \mid \text{No Disease})$

$\Rightarrow$ This is why we cannot estimate the relative risk from case-control data!

Odds Ratio Revisited

- The odds ratio measures association in Case-Control studies
- Odds $= \dfrac{\text{P(event occurs)}}{\text{P(event does not occur)}} = \frac{p}{1-p}$
- **Remember** the odds ratio is simply a ratio of odds!
- $OR = \dfrac{\text{odds in group 1}}{\text{odds in group 2}}$
- $OR = \frac{\hat{p}_1/(1-\hat{p}_1)}{\hat{p}_2/(1-\hat{p}_2)}$

Defining p_1 and p_2

- We can actually calculate OR using either "case-control" or "cohort" set up
- Using "case-control" p_1 and p_2 where we condition on disease or no disease

$$OR = \frac{(46/120)/(74/120)}{(12048/23562)/(11514/23562)} = \frac{46/74}{12048/11514} = 0.59$$

- Using "cohort" p_1 and p_2 where we condition on exposure or no exposure

$$OR = \frac{(46/12094)/(12048/12094)}{(74/11588)/(11514/11588)} = \frac{46/12048}{74/11514} = 0.59$$

- We get the same answer either way!

Confidence Interval for Odds Ratio

- Step 1: Find the estimate of the log OR

$$\log\left(\frac{\hat{p}_1/(1-\hat{p}_1)}{\hat{p}_2/(1-\hat{p}_2)}\right)$$

- Step 2: Estimate the variance of the log(OR) as:

$$\frac{1}{n_1p_1}+\frac{1}{n_1q_1}+\frac{1}{n_2p_2}+\frac{1}{n_2q_2}$$

- Step 3: Find the 95% CI for log(OR):

$$\log(OR) \pm 1.96 \cdot SD(\log OR) = (\text{lower, upper})$$

- Step 4: Exponentiate to get 95% CI for OR;

$$e^{(\text{lower, upper})}$$

Confidence Interval for Odds Ratio

Can we adopt the same reasoning as for the RR when computing the variance of the OR?

RR and OR

In Summary

- The relative risk cannot be estimated from a case-control study
- The odds ratio can be estimated from a case-control study
- The OR estimates the RR when the disease is rare in both groups:

$$OR = \frac{P(diseased|exposed)/(1 - P(diseased|exposed))}{P(diseased|unexposed)/(1 - P(diseased|unexposed)},$$

$$RR = \frac{P(diseased|exposed)}{P(diseased|unexposed)}$$

- The OR is invariant to cohort or case-control designs, the RR is not
- The OR is an essential concept in “logistic regression”, which is a generalization of “regression”