Show me your proof: Confidence intervals and p-values

Steve Simon

P.Mean Consulting

www.pmean.com

2. Why do I offer this webinar for free?

I offer free statistics webinars partly for fun and partly to build up goodwill for my consulting business,

- www.pmean.com/consult.html.

I also provide a free newsletter about Statistics, The Monthly Mean. To sign up for the newsletter, go to

- www.pmean.com/news

3. Abstract

- P-values and confidence intervals are the fundamental tools used in most inferential data analyses. They are possibly the most commonly reported statistics in the medical literature. Unfortunately, both p-values and confidence intervals are subject to frequent misinterpretations.

4. Abstract

- In this presentation, you will learn the proper interpretation of p-values and confidence intervals, and the common abuses and misconceptions about these statistics. You will also see a simple application of Bayesian analysis which provides an alternative to p-values and confidence intervals.

5. Learning objectives

- In this seminar, you will learn how to:
  - distinguish between statistical significance and clinical significance;
  - define and interpret p-values;
  - explain the ethical issues associated with inadequate sample sizes.

6. Outline

1. Pop quiz
2. Definitions
3. What is a p-value?
4. Practice exercises
5. What is a confidence interval?
6. Practice exercises
7. Repeat of pop quiz

7. Pop quiz

A research paper computes a p-value of 0.45. How would you interpret this p-value?

1. Strong evidence for the null hypothesis
2. Strong evidence for the alternative hypothesis
3. Little or no evidence for the null hypothesis
4. Little or no evidence for the alternative hypothesis
5. More than one answer above is correct.
6. I do not know the answer.

8. Pop quiz

A research paper computes a confidence interval for a relative risk of 0.82 to 3.94. What does this confidence interval tell you.

1. The result is statistically significant and clinically important.
2. The result is not statistically significant, but is clinically important.
3. The result is statistically significant, but not clinically important.
4. The result is not statistically significant, and not clinically important.
5. The result is ambiguous.
6. I do not know the answer.

9. Definitions: Population

- A population is a collection of items of interest in research. The population represents a group that you wish to generalize your research to. Populations are often **defined in terms of demography, geography, occupation, time, care requirements, diagnosis, or some combination of the above**.

10. Definitions: Population

- A population is a collection of items of interest in research. The population represents a group that you wish to generalize your research to. Populations are often **defined in terms of demography, geography, occupation, time, care requirements, diagnosis, or some combination of the above**.

11. Definitions: Population

- An example of a population would be all infants born in the state of Missouri during the 1995 calendar year who have one or more visits to the Emergency room during their first year of life.

12. Definitions: Sample

- A sample is a subset of a population. A random sample is a subset where every item in the population has the same probability of being in the sample. **Usually, the size of the sample is much less than the size of the population**. The primary goal of much research is to use information collected from a sample to try to characterize a certain population.

13. Definitions: Type I Error

- In your research, you specify a null hypothesis (typically labeled H0) and an alternative hypothesis (typically labeled Ha, or sometimes H1). By tradition, the null hypothesis corresponds to no change. A Type I error is **rejecting the null hypothesis when the null hypothesis is true**.

14. Definitions: Type I Error

- **Example:** Consider a new drug that we will put on the market if we can show that it is better than a placebo. In this context, H0 would represent the hypothesis that the average improvement (or perhaps the probability of improvement) among all patients taking the new drug is equal to the average improvement (probability of improvement) among all patients taking the placebo. A Type I error would be **allowing an ineffective drug onto the market**.

15. Definitions: Type II Error

- A Type II error is **accepting the null hypothesis when the null hypothesis is false**. **Many studies have small sample sizes that make it difficult to reject the null hypothesis,** even when there is a big change in the data. In these situations, a Type II error might be a possible explanation for the negative study results.

16. Definitions: Type II Error

- **Example:** Consider a new drug that we will put on the market if we can show that it is better than a placebo. In this context, H0 would represent the hypothesis that the average improvement (or perhaps the probability of improvement) among all patients taking the new drug is equal to the average improvement (probability of improvement) among all patients taking the placebo. A Type II error would be **keeping an effective drug off the market**.

17. What is a p-value?

- A p-value is a **measure of how much evidence we have against the null hypothesis**. The null hypothesis, traditionally represented by the symbol H0, represents the hypothesis of no change or no effect. The smaller the p-value, the more evidence we have against H0.

18. What is a p-value?

- The p-value is also a measure of how likely we are to get a certain sample result or a result “more extreme,” assuming H0 is true. The type of hypothesis (right tailed, left tailed or two tailed) will determine what “more extreme” means.

19. What is a p-value?

- The p-value is also a measure of how likely we are to get a certain sample result or a result “more extreme,” assuming H0 is true. The type of hypothesis (right tailed, left tailed or two tailed) will determine what “more extreme” means.

20. What is a p-value?

- It is easiest to understand the p-value in a data set that is already at an extreme. Suppose that a drug company alleges that **only 50% of all patients who take a certain drug will have an adverse event of some kind**. You believe that the adverse event rate is much higher. **In a sample of 12 patients, all twelve have an adverse event**.

21. What is a p-value?

- **The data supports your belief because it is inconsistent with the assumption of a 50% adverse event rate**. It would be like flipping a coin 12 times and getting heads each time.

22. What is a p-value?

- The p-value, the probability of getting a sample result of **12 adverse events in 12 patients** assuming that the adverse event rate is 50%, is a measure of this inconsistency. **The p-value, 0.000244, is small enough that we would reject the hypothesis that the adverse event rate was only 50%.**

23. What is a p-value?

A large p-value should not automatically be construed as evidence in support of the null hypothesis. Perhaps the failure to reject the null hypothesis was caused by an inadequate sample size. When you see a large p-value in a research study, you should also look for one of two things:

1. a **power calculation** that confirms that the sample size in that study was adequate for detecting a clinically relevant difference; and/or
2. a **confidence interval** that lies entirely within the range of clinical indifference.

24. What is a p-value?

You should also be cautious about a small p-value, but for different reasons. **In some situations, the sample size is so large that even differences that are trivial from a medical perspective can still achieve statistical significance**.

25. Practice exercise: interpret the p-values shown below.

1. **The Outcome of Extubation Failure in a Community Hospital Intensive Care Unit: A Cohort Study**. Seymour CW, Martinez A, Christie JD, Fuchs BD. *Critical Care 2004*, 8:R322-R327 (20 July 2004) **Introduction:** Extubation failure has been associated with poor intensive care unit (ICU) and hospital outcomes in tertiary care medical centers. Given the large proportion of critical care delivered in the community setting, our purpose was to determine the impact of extubation failure on patient outcomes in a community hospital ICU. **Methods:** A retrospective cohort study was performed using data gathered in a 16-bed medical/surgical ICU in a community hospital. During 30 months, all patients with acute respiratory failure admitted to the ICU were included in the source population if they were mechanically ventilated by endotracheal tube for more than 12 hours. Extubation failure was defined as reinstitution of mechanical ventilation within 72 hours (n = 60), and the control cohort included patients who were successfully extubated at 72 hours (n = 93). **Results:** The primary outcome was total ICU length of stay after the initial extubation. Secondary outcomes were total hospital length of stay after the initial extubation, ICU mortality, hospital mortality, and total hospital cost. Patient groups were similar in terms of age, sex, and severity of illness, as assessed using admission Acute Physiology and Chronic Health Evaluation II score ($P > 0.05$). Both ICU (1.0 versus 10 days; $P < 0.01$) and hospital length of stay (6.0 versus 17 days; $P < 0.01$) after initial extubation were significantly longer in reintubated patients. ICU mortality was significantly higher in patients who failed extubation (odds ratio = 12.2, 95% confidence interval [CI] = 1.5–101; $P < 0.05$), but there was no significant difference in hospital mortality (odds ratio = 2.1, 95% CI = 0.8–5.4; $P < 0.15$). Total hospital costs (estimated from direct and indirect charges) were significantly increased by a mean of US$33,926 (95% CI = US$22,573–45,280; $P < 0.01$). **Conclusion:** Extubation failure in a community hospital is univariately associated with prolonged inpatient care and significantly increased cost. Corroborating data from tertiary care centers, these adverse outcomes highlight the importance of accurate predictors of extubation outcome.

26. Practice exercise: interpret the p-values shown below.

2. **Elevated White Cell Count in Acute Coronary Syndromes: Relationship to Variants in Inflammatory and Thrombotic Genes**. Byrne CE, Fitzgerald A, Cannon CP, Fitzgerald DJ, Shields DC. *BMC Medical Genetics 2004*, 5:13 (1 June 2004) **Background:** Elevated white blood cell counts (WBC) in acute coronary syndromes (ACS) increase the risk of recurrent events, but it is not known if this is exacerbated by pro-inflammatory factors. We sought to identify whether pro-inflammatory genetic variants contributed to alterations in WBC and C-reactive protein (CRP) in an ACS population. **Methods:** WBC and genotype of interleukin 6 (IL-6 G-174C) and of interleukin-1 receptor antagonist (IL1RN intronic repeat polymorphism) were investigated in 732 Caucasian patients with ACS in the OPUS-TIMI-16 trial. Samples for measurement of WBC and inflammatory factors were taken at baseline, i.e. Within 72 hours of an acute myocardial infarction or an unstable angina event. **Results:** An increased white blood cell count (WBC) was associated with an increased C-reactive protein ($r = 0.23$, $p < 0.001$) and there was also a positive correlation between levels of β-fibrinogen and C-reactive protein ($r = 0.42$, $p < 0.0001$). IL1RN and IL6 genotypes had no significant impact upon WBC. The difference in median WBC between the two homozygote IL6 genotypes was 0.21/mm3 (95% CI = -0.41, 0.77), and -0.03/mm3 (95% CI = -0.55, 0.86) for IL1RN. Moreover, the composite endpoint was not significantly affected by an interaction between WBC and the IL1 ($p = 0.61$) or IL6 ($p = 0.48$) genotype. **Conclusions:** Cytokine pro-inflammatory genetic variants do not influence the increased inflammatory profile of ACS patients.

27. Practice exercise: interpret the p-values shown below.

3. **Is There a Clinically Significant Gender Bias in Post-Myocardial Infarction Pharmacological Management in the Older (>60) Population of a Primary Care Practice?** Di Cecco R, Patel U, Upshur REG. *BMC Family Practice 2002*, 3:8 (3 May 2002) **Background:** Differences in the management of coronary artery disease between men and women have been reported in the literature. There are few studies of potential inequalities of treatment that arise from a primary care context. This study investigated the existence of such inequalities in the medical management of post myocardial infarction in older patients. **Methods:** A comprehensive chart audit was conducted of 142 men and 81 women in an academic primary care practice. Variables were extracted on demographic variables, cardiovascular risk factors, medical and non-medical management of myocardial infarction. **Results:** Women were older than men. The groups were comparable in terms of cardiac risk factors. A statistically significant difference (14.6%: 95% CI 0.048–28.7 p = 0.047) was found between men and women for the prescription of lipid lowering medications. 25.3% (p = 0.0005, CI 11.45, 39.65) more men than women had undergone angiography, and 14.4 % (p = 0.029, CI 2.2, 26.6) more men than women had undergone coronary artery bypass graft surgery. **Conclusion:** Women are less likely than men to receive lipid-lowering medication which may indicate less aggressive secondary prevention in the primary care setting.

28. **What is a confidence interval?**

- We statisticians have a habit of **hedging our bets**. We always insert qualifiers into our reports, warn about all sorts of assumptions, and never admit to anything more extreme than probable. There's a famous saying: **"Statistics means never having to say you're certain."**

29. **What is a confidence interval?**

- We qualify our statements, of course, because we are always **dealing with imperfect information**. In particular, we are often asked to make statements about a population (a large group of subjects) using information from a sample (a small, but carefully selected subset of this population). No matter how carefully this sample is selected to be a fair and unbiased representation of the population, **relying on information from a sample will always lead to some level of uncertainty**.

30. **What is a confidence interval?**

- **A confidence interval is a range of values that tries to quantify uncertainty associated with the sampling process**. Consider it as a **range of plausible values**.

31. **What is a confidence interval?**

- A wide confidence interval implies poor precision; we can only specify plausible values to a broad and uninformative range. A narrow confidence interval implies good precision; we can place sharp limits on the range of plausible values.

32. **What is a confidence interval?**

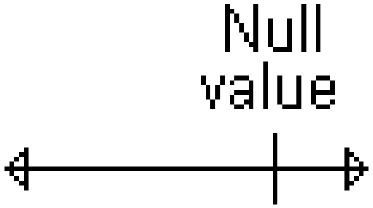
- Consider a recent study of **homoeopathic treatment of pain and swelling after oral surgery** (Lokken 1995). When examining swelling 3 days after the operation, they showed that **homoeopathy led to 1 mm less swelling on average**. The **95% confidence interval, however, ranged from -5.5 to 7.5 mm**. This interval implies that **neither a large improvement due to homoeopathy nor a large decrement could be ruled out**.

33. **What is a confidence interval?**

- When you see a confidence interval in a published medical report, you should look for two things. First, **does the interval contain a value that implies no change or no effect**? For example, with a confidence interval for a difference look to see whether that interval includes zero. With a confidence interval for a ratio, look to see whether that interval contains one.

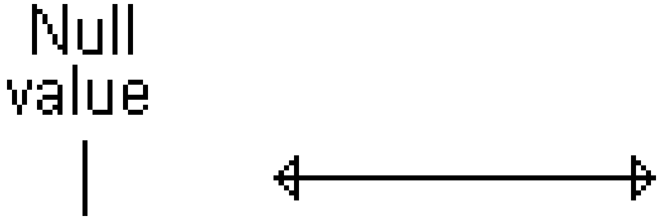
34. **What is a confidence interval?**

- Here's an example of a confidence interval that contains the null value. **The interval shown below implies no statistically significant change**.



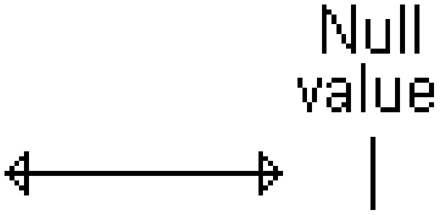
35. **What is a confidence interval?**

- Here's an example of a confidence interval that excludes the null value. If we assume that larger implies better, then **the interval shown below would imply a statistically significant improvement**.



36. **What is a confidence interval?**

- Here's a different example of a confidence interval that excludes the null value. **The interval shown below implies a statistically significant decline**.

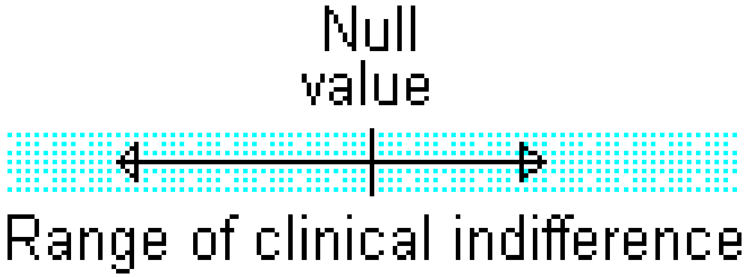


37. **What is a confidence interval?**

- You should also see **whether the confidence interval lies partly or entirely within a range of clinical indifference**. Clinical indifference represents values of such a trivial size that you would not want to change your current practice.

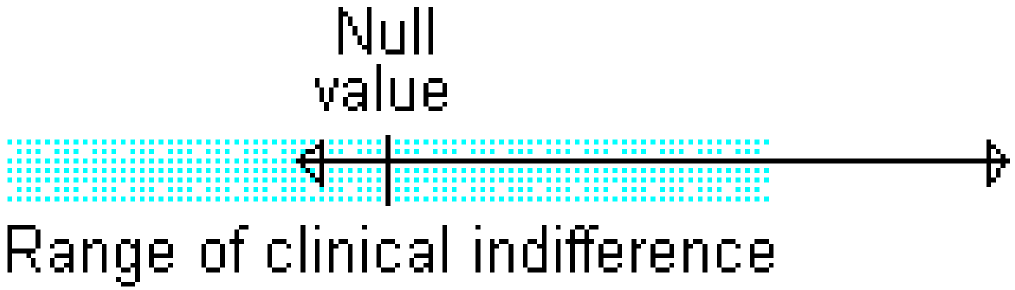
38. **What is a confidence interval?**

- **If a confidence interval is contained entirely within your range of clinical indifference**, then you have clear and convincing evidence to **keep doing things the same way**.



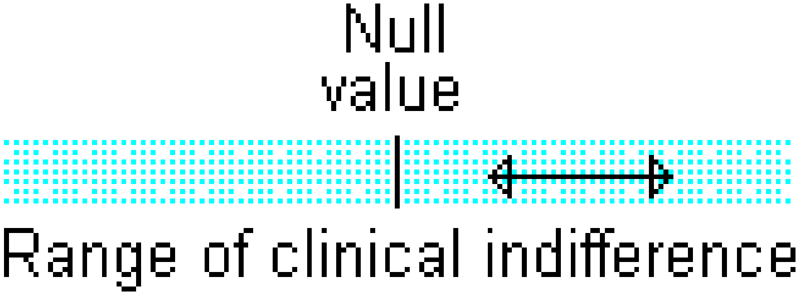
39. **What is a confidence interval?**

- One the other hand, **if part of the confidence interval lies outside the range of clinical indifference**, then you should consider the possibility that **the sample size is too small**.



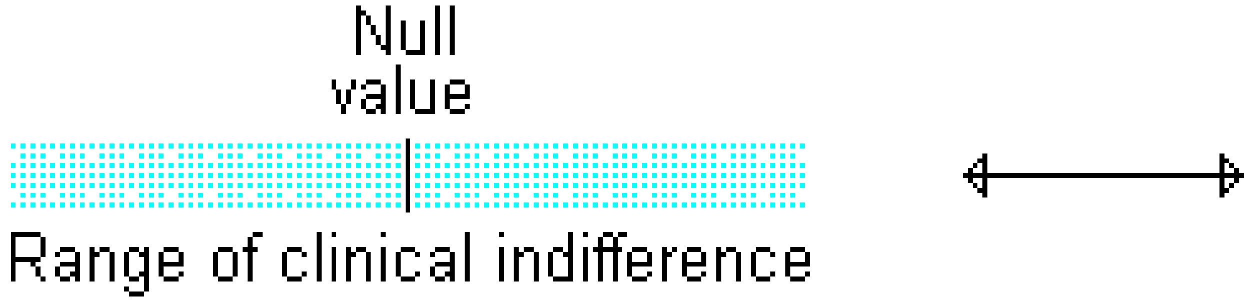
40. **What is a confidence interval?**

- If your **confidence interval excludes the null value but still lies entirely within the range of clinical indifference**, then you have a result with **statistical significance, but no practical significance**.



41. **What is a confidence interval?**

- Finally, if your **confidence interval excludes the null value and lies outside the range of clinical indifference**, then you have **both statistical and practical significance**.



42. Practice exercise: interpret the confidence intervals shown below.

1. **The Outcome of Extubation Failure in a Community Hospital Intensive Care Unit: A Cohort Study**. Seymour CW, Martinez A, Christie JD, Fuchs BD. *Critical Care 2004*, 8:R322-R327 (20 July 2004) **Introduction:** Extubation failure has been associated with poor intensive care unit (ICU) and hospital outcomes in tertiary care medical centers. Given the large proportion of critical care delivered in the community setting, our purpose was to determine the impact of extubation failure on patient outcomes in a community hospital ICU. **Methods:** A retrospective cohort study was performed using data gathered in a 16-bed medical/surgical ICU in a community hospital. During 30 months, all patients with acute respiratory failure admitted to the ICU were included in the source population if they were mechanically ventilated by endotracheal tube for more than 12 hours. Extubation failure was defined as reinstitution of mechanical ventilation within 72 hours (n = 60), and the control cohort included patients who were successfully extubated at 72 hours (n = 93). **Results:** The primary outcome was total ICU length of stay after the initial extubation. Secondary outcomes were total hospital length of stay after the initial extubation, ICU mortality, hospital mortality, and total hospital cost. Patient groups were similar in terms of age, sex, and severity of illness, as assessed using admission Acute Physiology and Chronic Health Evaluation II score ($P > 0.05$). Both ICU (1.0 versus 10 days; $P < 0.01$) and hospital length of stay (6.0 versus 17 days; $P < 0.01$) after initial extubation were significantly longer in reintubated patients. ICU mortality was significantly higher in patients who failed extubation (odds ratio = 12.2, 95% confidence interval [CI] = 1.5–101; $P < 0.05$), but there was no significant difference in hospital mortality (odds ratio = 2.1, 95% CI = 0.8–5.4; $P < 0.15$). Total hospital costs (estimated from direct and indirect charges) were significantly increased by a mean of US$33,926 (95% CI = US$22,573–45,280; $P < 0.01$). **Conclusion:** Extubation failure in a community hospital is univariately associated with prolonged inpatient care and significantly increased cost. Corroborating data from tertiary care centers, these adverse outcomes highlight the importance of accurate predictors of extubation outcome.

43. Practice exercise: interpret the confidence intervals shown below.

2. **Elevated White Cell Count in Acute Coronary Syndromes: Relationship to Variants in Inflammatory and Thrombotic Genes**. Byrne CE, Fitzgerald A, Cannon CP, Fitzgerald DJ, Shields DC. *BMC Medical Genetics 2004*, 5:13 (1 June 2004) **Background:** Elevated white blood cell counts (WBC) in acute coronary syndromes (ACS) increase the risk of recurrent events, but it is not known if this is exacerbated by pro-inflammatory factors. We sought to identify whether pro-inflammatory genetic variants contributed to alterations in WBC and C-reactive protein (CRP) in an ACS population. **Methods:** WBC and genotype of interleukin 6 (IL-6 G-174C) and of interleukin-1 receptor antagonist (IL1RN intronic repeat polymorphism) were investigated in 732 Caucasian patients with ACS in the OPUS-TIMI-16 trial. Samples for measurement of WBC and inflammatory factors were taken at baseline, i.e. Within 72 hours of an acute myocardial infarction or an unstable angina event. **Results:** An increased white blood cell count (WBC) was associated with an increased C-reactive protein ($r = 0.23$, $p < 0.001$) and there was also a positive correlation between levels of β-fibrinogen and C-reactive protein ($r = 0.42$, $p < 0.0001$). IL1RN and IL6 genotypes had no significant impact upon WBC. The difference in median WBC between the two homozygote IL6 genotypes was 0.21/mm3 (95% CI = -0.41, 0.77), and -0.03/mm3 (95% CI = -0.55, 0.86) for IL1RN. Moreover, the composite endpoint was not significantly affected by an interaction between WBC and the IL1 ($p = 0.61$) or IL6 ($p = 0.48$) genotype. **Conclusions:** Cytokine pro-inflammatory genetic variants do not influence the increased inflammatory profile of ACS patients.

44. Practice exercise: interpret the confidence intervals shown below.

3. **Is There a Clinically Significant Gender Bias in Post-Myocardial Infarction Pharmacological Management in the Older (>60) Population of a Primary Care Practice?** Di Cecco R, Patel U, Upshur REG. *BMC Family Practice 2002*, 3:8 (3 May 2002) **Background:** Differences in the management of coronary artery disease between men and women have been reported in the literature. There are few studies of potential inequalities of treatment that arise from a primary care context. This study investigated the existence of such inequalities in the medical management of post myocardial infarction in older patients. **Methods:** A comprehensive chart audit was conducted of 142 men and 81 women in an academic primary care practice. Variables were extracted on demographic variables, cardiovascular risk factors, medical and non-medical management of myocardial infarction. **Results:** Women were older than men. The groups were comparable in terms of cardiac risk factors. A statistically significant difference (14.6%: 95% CI 0.048–28.7 p = 0.047) was found between men and women for the prescription of lipid lowering medications. 25.3% (p = 0.0005, CI 11.45, 39.65) more men than women had undergone angiography, and 14.4 % (p = 0.029, CI 2.2, 26.6) more men than women had undergone coronary artery bypass graft surgery. **Conclusion:** Women are less likely than men to receive lipid-lowering medication which may indicate less aggressive secondary prevention in the primary care setting.

45. Repeat of pop quiz

A research paper computes a p-value of 0.45. How would you interpret this p-value?

1. Strong evidence for the null hypothesis
2. Strong evidence for the alternative hypothesis
3. Little or no evidence for the null hypothesis
4. Little or no evidence for the alternative hypothesis
5. More than one answer above is correct.
6. I do not know the answer.

46. Repeat of pop quiz

A research paper computes a confidence interval for a relative risk of 0.82 to 3.94. What does this confidence interval tell you.

1. The result is statistically significant and clinically important.
2. The result is not statistically significant, but is clinically important.
3. The result is statistically significant, but not clinically important.
4. The result is not statistically significant, and not clinically important.
5. The result is ambiguous.
6. I do not know the answer.