Statistical Inference

Last modified: 2026-09-29 00:05:17 (PDT)

Inference

Definition 0.1 (Statistical inference)**. Statistical inference** is the process of analyzing data to learn about the shape and structure of the probability distribution that generated those data.

This definition is adapted from Wikipedia's (Wikipedia contributors 2025).

Example 0.1 (Inference about cholesterol in the WCGS population)**.** The WCGS measured total cholesterol in 3,142 of its 3,154 participants. Using those measurements to learn about the distribution of cholesterol among men like the participants, for example its mean, and quantifying how uncertain that knowledge is, is statistical inference.

Statistical inference typically consists of two steps:

1. fitting a statistical model to data;
2. summarizing our uncertainty about the parameters of the fitted model, based on the data (and, in Bayesian inference, on prior beliefs).

There are two predominant paradigms for statistical inference:

- frequentist inference, which treats parameters as fixed unknown constants;
- Bayesian inference, which treats parameters as random variables with prior distributions.

Hypothesis tests

Hypotheses

Definition 0.2 (Null hypothesis)**.** The **null hypothesis**, written H_0, is a specific claim about the parameters of a statistical model, which a hypothesis test assesses against the data.

Definition 0.3 (Alternative hypothesis)**.** The **alternative hypothesis**, written H_1 or H_A, is the claim that a hypothesis test weighs the data as evidence *for*, against the null hypothesis.

Definition 0.4 (Two-sided alternative)**.** For a null hypothesis that a parameter equals a value, such as $H_0 : \theta = \theta_0$, the **two-sided alternative** is that the parameter differs from that value in either direction: $H_1 : \theta \neq \theta_0$.

Example 0.2 (Cholesterol and CHD in the WCGS)**.** Let μ_1 and μ_0 be the mean cholesterol of men who do and do not have a CHD event by 1969, in the population the WCGS sampled. A comparison of the two groups tests the null hypothesis that the means are equal against the two-sided alternative that they differ:

$$H_0 : \mu_1 = \mu_0 \qquad H_1 : \mu_1 \neq \mu_0$$

p-values

Definition 0.5 (Test statistic)**.** A **test statistic** T is a function of the data whose distribution under the null hypothesis is known, exactly or approximately, and whose larger values are more extreme, meaning less consistent with H_0 and more consistent with H_1.

Definition 0.6 (p-value)**.** Given a test statistic T whose observed value is t, the **p-value** is the probability that T is at least as extreme as the value observed, computed with the data distributed as the null hypothesis specifies:

$$p \stackrel{\text{def}}{=} \Pr_{H_0}(T \geq t)$$

If the null hypothesis allows more than one distribution (a *composite* null hypothesis, such as $H_0 : \mu_1 = \mu_0$ with unknown variances), the p-value is the largest such probability over those distributions, or an approximation to it.

A p-value is computed assuming H_0 is true, so it is not the probability that H_0 is true given the data.

Definition 0.7 (Significance level)**.** A hypothesis test with **significance level** α rejects the null hypothesis when the p-value is at most α.

$\alpha = 0.05$ is a common convention, not a rule. When H_0 is true and the p-value is computed exactly, a test with significance level α rejects H_0 with probability at most α; for a test based on an approximate p-value, such as a large-sample test, this holds only approximately.

Definition 0.8 (Statistically significant)**.** A result is **statistically significant** at level α if it leads a test with significance level α to reject the null hypothesis.

Example 0.3 (Testing whether mean cholesterol differs by CHD status)**.** Continuing Example 0.2, let $\bar{x}_1$ and $\bar{x}_0$ be the two groups' sample mean cholesterol, with sample variances s_1^2 and s_0^2 and sample sizes n_1 and n_0. By the central limit theorem, when H_0 holds and the samples are large, the statistic

$$Z \stackrel{\text{def}}{=} \frac{\bar{x}_1 - \bar{x}_0}{\sqrt{s_1^2/n_1 + s_0^2/n_0}}$$

has approximately a standard Gaussian distribution. Extreme values in either direction are evidence for the two-sided H_1, so the test statistic is $T = |Z|$, and the p-value is $\Pr(|Z| \geq |z|) = 2\Phi(-|z|)$, where Φ is the standard Gaussian CDF.

```r
chol_by_chd <- split(wcgs$chol, wcgs$chd69) |>
  lapply(function(x) x[!is.na(x)])
x1 <- chol_by_chd[["Yes"]]
x0 <- chol_by_chd[["No"]]
diff_means <- mean(x1) - mean(x0)
se_diff <- sqrt(var(x1) / length(x1) + var(x0) / length(x0))
z <- diff_means / se_diff
p_value <- 2 * pnorm(-abs(z))
c(difference = diff_means, se = se_diff, z = z, p_value = p_value)
#>  difference          se           z     p_value
#> 2.58087e+01 3.17988e+00 8.11626e+00 4.80780e-16
```

The p-value is far below 0.05, so at significance level 0.05 the test rejects H_0: the data provide statistically significant evidence that mean cholesterol differs between the two groups.

Reference distributions

The tests on this page compare a test statistic with one of three families of distributions, each built from independent standard Gaussian random variables. Here, "independent" means mutually independent.

Definition 0.9 (Chi-square distribution)**.** Let $Z_1, \ldots, Z_k$ be independent random variables, each with the standard Gaussian distribution. The distribution of $\sum_{j=1}^{k} Z_j^2$ is the **chi-square distribution with k degrees of freedom**, written χ_k^2.

Example 0.4 (The 0.95 quantile of χ_1^2)**.** By Definition 0.9, a χ_1^2 random variable is Z^2 for a standard Gaussian Z. So $\Pr(Z^2 \leq c) = \Pr(-\sqrt{c} \leq Z \leq \sqrt{c})$, and the 0.95 quantile of χ_1^2 is the square of the 0.975 quantile of the standard Gaussian distribution:

```
c(chisq = qchisq(0.95, df = 1), gaussian_squared = qnorm(0.975)^2)
#>            chisq gaussian_squared
#>          3.84146          3.84146
```

Definition 0.10 (t-distribution)**.** Let Z have the standard Gaussian distribution, let V have the χ_k^2 distribution (Definition 0.9), and let Z and V be independent. The distribution of

$$\frac{Z}{\sqrt{V/k}}$$

is the **t-distribution with k degrees of freedom** (or Student's t-distribution), written t_k.

Example 0.5 (t quantiles approach Gaussian quantiles)**.** The 0.975 quantile of t_k is larger than the standard Gaussian's 0.975 quantile, 1.96, and approaches it as k grows:

```r
k <- c(4, 9, 29, 99, 2760)
tibble::tibble(k = k, t_quantile = qt(0.975, df = k))
#> # A tibble: 5 x 2
#>       k t_quantile
#>   <dbl>      <dbl>
#> 1     4       2.78
#> 2     9       2.26
#> 3    29       2.05
#> 4    99       1.98
#> 5  2760       1.96
```

So t-based intervals and tests differ noticeably from Gaussian-based ones only in small samples.

Definition 0.11 (F-distribution)**.** Let U have the $\chi^2_{k_1}$ distribution, let V have the $\chi^2_{k_2}$ distribution, and let U and V be independent. The distribution of

$$\frac{U/k_1}{V/k_2}$$

is the **F-distribution with k_1 and k_2 degrees of freedom**, written F_{k_1,k_2}.

Example 0.6 (A squared t random variable has an F-distribution)**.** Let $T = Z/\sqrt{V/k}$ as in Definition 0.10. Then

$$T^2 = \frac{Z^2/1}{V/k},$$

where Z^2 has the χ^2_1 distribution (Definition 0.9) and is independent of V. So T^2 has the $F_{1,k}$ distribution (Definition 0.11), and the 0.95 quantile of $F_{1,k}$ is the square of the 0.975 quantile of t_k:

```r
c(f = qf(0.95, df1 = 1, df2 = 9), t_squared = qt(0.975, df = 9)^2)
#>         f t_squared
#>   5.11736   5.11736
```

Confidence intervals

Definition 0.12 (Confidence interval)**.** A $100(1-\alpha)\%$ **confidence interval** for a parameter θ is a pair of statistics L and U, computed from the data, such that for every possible value of θ:

$$\Pr(L \le \theta \le U) = 1 - \alpha$$

The probability is over repeated samples from the data-generating process, with θ held fixed.

Definition 0.13 (Coverage probability)**.** The **coverage probability** of an interval $[L, U]$ for a parameter θ is $\Pr(L \le \theta \le U)$, computed over repeated samples with θ held fixed. A $100(1-\alpha)\%$ confidence interval has coverage probability $1-\alpha$.

Definition 0.14 (Approximate confidence interval)**.** An **approximate** $100(1-\alpha)\%$ confidence interval is an interval whose coverage probability is approximately $1-\alpha$, for example only in large samples.

Example 0.7 (Confidence interval for mean cholesterol in the WCGS)**.** By the central limit theorem, the sample mean $\bar{X}$ of a large sample has approximately a Gaussian distribution with mean μ and standard error $\sigma/\sqrt{n}$, which we estimate by $s/\sqrt{n}$. So $\bar{x} \pm z_{0.975}\, s/\sqrt{n}$, where $z_{0.975} \approx 1.96$ is the 0.975 quantile of the standard Gaussian distribution, is an approximate 95% confidence interval for μ:

```
chol <- wcgs$chol[!is.na(wcgs$chol)]
se_chol <- sd(chol) / sqrt(length(chol))
mean(chol) + c(lower = -1, upper = 1) * qnorm(0.975) * se_chol
#>   lower   upper
#> 224.854 227.891
```

Definition 0.15 (Margin of error)**.** The **margin of error** (or **radius**) of a confidence interval is half the interval's width.

Example 0.8 (Margin of error for mean cholesterol)**.** In Example 0.7, the margin of error is $z_{0.975} \times s/\sqrt{n} \approx 1.52$ mg/dL.

Further reading on confidence intervals

For more on confidence intervals:

- Anatomy of a confidence interval (PDF);
- https://www.youtube.com/watch?v=vq1KrE7gU5M

Interpretation of negative findings

If a confidence interval includes the null value, or a hypothesis test fails to reject the null hypothesis, that result does not *necessarily* mean that the null hypothesis is true. (When the interval and the test are built from the same statistic, with coverage $1 - \alpha$ and significance level α, the two results coincide.) So we should not interpret such results as "the odds (or risks, hazards, or means) are not significantly different". Instead, we should write something like "the data do not provide statistically significant *evidence* that the odds (or risks, hazards, or means) differ". Statistical significance is a property of evidence, not of the estimands.

P-values do not distinguish between absence of evidence and evidence of absence. Confidence intervals do:

- a narrow confidence interval that includes the null value is evidence of absence (of any effect large enough to matter);
- a confidence interval that includes the null value but also includes substantially non-null values represents absence of evidence.

Conversely, even statistically significant evidence of a non-null value does not mean the value is large enough to matter; that depends on what the estimand is. For example, we might have statistically significant evidence that an exercise program prolongs human lifespans by 20 seconds, but an effect that small would be negligible in practical terms.

Figure 1 sketches these scenarios.

See also (Vittinghoff et al. 2012, sec. 3.7).

References

Vittinghoff, Eric, David V Glidden, Stephen C Shiboski, and Charles E McCulloch. 2012. *Regression Methods in Biostatistics: Linear, Logistic, Survival, and Repeated Measures Models*. 2nd ed. Springer. https://doi.org/10.1007/978-1-4614-1353-0.

Wikipedia contributors. 2025. *Statistical Inference — Wikipedia, the Free Encyclopedia*. https://en.wikipedia.org/w/index.php?title=Statistical_inference&oldid=1304071803.

```r
ci_scenarios <- tibble::tribble(
  ~scenario, ~lower, ~upper,
  "not enough data: we know almost nothing", 0.2, 5,
  "absence of evidence", 0.5, 2,
  "\"trending toward significance\" (still absence of evidence)", 0.95, 2.5,
  "evidence of absence", 0.97, 1.03,
  "evidence of a negligible effect", 1.02, 1.1,
  "uncertain whether negligible or substantial", 1.05, 1.8,
  "evidence of a substantial effect", 1.7, 1.8,
  "substantial, but imprecise", 1.5, 4
) |>
  dplyr::mutate(scenario = factor(scenario, levels = rev(scenario)))

ci_scenarios |>
  ggplot2::ggplot() +
  ggplot2::aes(y = scenario, xmin = lower, xmax = upper) +
  ggplot2::geom_errorbarh(height = 0.3) +
  ggplot2::geom_vline(xintercept = 1, linetype = "dashed") +
  ggplot2::scale_x_log10() +
  ggplot2::labs(x = "Ratio estimand (log scale); null value = 1", y = NULL)
```

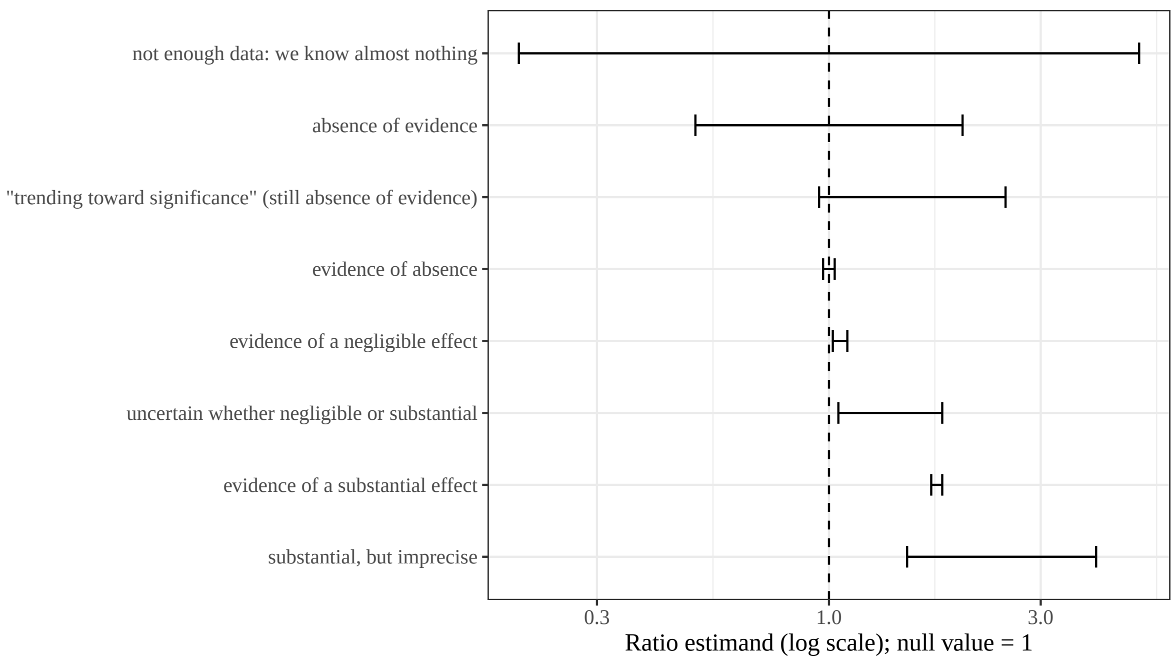

Figure 1: Interpretations of hypothetical 95% confidence intervals for a ratio (such as an odds ratio), relative to the null value 1 (dashed line). Adapted from a whiteboard sketch made in office hours.