Mayo's Post-data Severity Evaluation

Statistik Kolloquium

André Meichtry

March 5, 2024

Data and statistical Test

```
mu0 <- 0  #H0
sigma <- 2  #known SD
n <- 30  #sample size
alpha <- 0.025  #siglevel
crit <- mu0 + qnorm(1 - alpha) * sigma/sqrt(n)  # critical value
d <- crit + 0.01  #observed distance from H0
xbar <- mu0 + d  #observed mean
t <- (xbar - mu0)/(sigma/sqrt(n))  #observed t-statistic
p <- 1 - pnorm(t)  #p-value
```

Consider the test T of $H_0 : \mu \leq \mu_0$ versus $H_1 : \mu > \mu_0$ with $\alpha = 0.025$, $\mu_0 = 0$, $n = 30$, assume $\sigma = 2$ known.

The common rule is: Reject H_0, if $\bar{X} > \bar{x}_{crit} = \mu_0 + z_{1-\alpha} \cdot \sigma/\sqrt{n} = 0.7157$ or, equivalently, if $Z = \frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}} > z_{1-\alpha} = 1.96$.

Assume now the observed data is $\bar{x} = 0.7257$. That is, we **reject** $H_0 : \mu_0 \leq 0$. The one-sample z-test gives $z = 1.9874$ and $p = 0.0234$.

Post-data Severity Evaluation

Deborah Mayo: https://en.wikipedia.org/wiki/Deborah_Mayo

Definition

Assume a **claim** $C : \mu > \mu_1$ and a **counter-claim** $\neg C : \mu \leq \mu_1$.

The Severity with which claim C passes test T with outcome x is defined as the probability that test T would have produced a result that accords less well with C than x does, if $\neg C$ were true

shortly:

$$\text{Sev}(T, x, \mu > \mu_1) = \text{Pr}(X \leq x \mid \mu \leq \mu_1)$$

This probability should be high.

Or, equivalently, the probability that test T would have produced a result that accords better with C than x does, if $\neg C$ were true,

$$1 - \text{Sev}(T, x, \mu > \mu_1) = \text{Pr}(X > x \mid \mu \leq \mu_1)$$

should be small!

This is a form of a modus tollens argument with two premises and a conclusion:

- If $\neg C \rightarrow X < x$.
- $X > x$.
- Therefore, not $\neg C$.

Implementation

```
library(severity)
```

```
severity                package:severity                R Documentation

Mayo's _Post-data_ Severity Evaluation

Description:

     Computes severity at various discrepancies (from the null
     hypothesis) for the hypothesis test H_{0}: mu = mu_{0} vs H_{1}:
     mu > mu_{0}, where mu_{0} is the hypothesized value. Also plots
     both the severity curve(s) and the power curve on a single plot.
```

```
sev <- severity(mu0 = mu0, xbar = xbar, sigma = sigma, n = n, alpha = alpha)
abline(v = mu0, lty = 2)
abline(v = xbar, lty = 4)
```

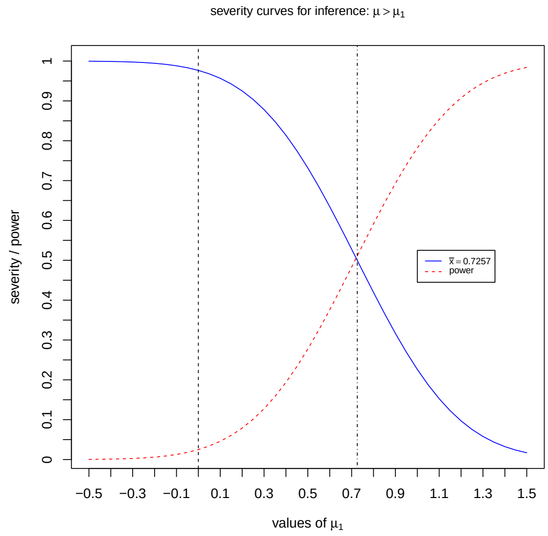

Figure 1: Severity and power curve for different claims after observing $\bar{x} = 0.726$ with test $\mu \leq 0$

Consider our actual test T of the null hypothesis $H_0 : \mu \leq 0$. We rejected H_0 after observing $\bar{x} = 0.7257 > \bar{x}_{crit} = 0.7157$ with $p = 0.0234$. The blue line represents the severity for different claims $\mu > \mu_1$ passing test T after observing $\bar{x} = 0.7257$. The red line is the power for different μ, the probability of rejecting H_0, if μ is true.

- the severity for the claim $\mu > 0$, the hypothesized null, is 0.9766.
- the severity for the claim $\mu > 0.7257$, the observed value, is 0.5.
- the severity for claims such as $\mu > 1.5$ is already very small, it is 0.017.
- If the observed value is around the critical value as in our case, $Power(\mu) = 1 - Severity(\mu)$.

Power vs. Severity

The alternative hypothesis $H_1 : \mu > \mu_0$ ($\mu > 0$ in our example) is a *composite* hypothesis. Assume that *prior* to a study, sample size n was calculated *assuming* a discrepancy from the null of $\mu = 1.5$. With $n = 30$ and $\sigma = 2$, this would lead to a power of 0.9841.

Researchers then say that they want to *detect* (better would be to say “signal”) an effect of $\mu \geq 1.5$ with high power, that is, the probability of rejecting $H_0 : \mu \leq 0$ should be 0.9841, **if $\mu = 1.5$ holds**.

It is now very important to note that *a posteriori*, with the data at hand ($\bar{x} = 0.7257$), we only reject $H_0 : \mu \leq 0$ in favor of $H_1 : \mu > 0$. The researcher has by no means “shown” that $\mu > 1.5$ holds.

Such statements are frequent. To claim that $\mu > 1.5$ after a significant test of $H_0 : \mu \leq 0$ has a very, very low severity.

The claim $\mu > 1.5$ is not severely tested, the severity for such a claim is only 0.017. The specified *simple* alternative $\mu = 1.5$ has **no role** a posteriori!

In science, theories and hypotheses that are not severely tested have – following Popper – a very low empirical content. We must say that – probably – in our field of research, a large number of theories and hypothesis are not severely tested. This is because our tests are often weak tests.

All analyses were performed using the R statistical software R version 4.3.3 (2024-02-29) [R C23].

References

[R C23] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria, 2023. URL: https://www.R-project.org/.

Session Info

- R version 4.3.3 (2024-02-29), `x86_64-pc-linux-gnu`
- Running under: `Ubuntu 22.04.4 LTS`
- Matrix products: default
- BLAS: `/usr/lib/x86_64-linux-gnu/blas/libblas.so.3.10.0`
- LAPACK: `/usr/lib/x86_64-linux-gnu/lapack/liblapack.so.3.10.0`
- Base packages: base, datasets, graphics, grDevices, methods, stats, utils
- Other packages: knitr 1.45, severity 2.0
- Loaded via a namespace (and not attached): compiler 4.3.3, evaluate 0.23, formatR 1.9, highr 0.10, tools 4.3.3, xfun 0.41