STAT 135, Concepts of Statistics

Helmut Pitters

Hypothesis testing 2

Department of Statistics
University of California, Berkeley

March 7, 2017

Hypothesis testing.

Example (Adjusting lab equipment)

Hospital lab uses instrument to determine hemoglobin levels in blood samples. Instrument needs to be calibrated on regular basis—many states require daily checks of instruments. To this end, lab makes measurements using sample from standard blood supply.

On particular day technician carries out n independent measurements of standard blood supply with known mean hemoglobin level 15.1.

$$H_0 \colon \mu = 15.1 \quad \text{(no adjustment needed)}$$
$$H_A \colon \mu \neq 15.1 \quad \text{(instrument needs to be readjusted)}$$

Hypothesis testing.

Example (Adjusting lab equipment)

Lab assumes $\bar{X}_n \sim \mathcal{N}(\mu, 0.16)$ and does not adjust instrument if error is within two standard deviations, i.e. has acceptance region

$$\bar{C} := 15.1 \pm 2 \times 0.4 = (14.3, 15.9)$$

yielding significance level

$$\begin{aligned}\alpha &= \mathbb{P}\{\text{type I error}\} = 1 - \mathbb{P}\left\{14.3 \leq \bar{X}_n \leq 15.9 | H_0\right\} \\ &= 1 - \left(\Phi(\frac{15.9 - 15.1}{0.4}) - \Phi(\frac{14.3 - 15.1}{0.4})\right) \approx 0.046\end{aligned}$$

Hypothesis testing.

Example (Adjusting lab equipment)

However, probability β of type II error is not uniquely determined, since alternative

$$H_A\colon \mu \neq 15.1$$

contains more than one distribution. For any particular value $\mu^\star \neq 15.1$ can find the power

$$\begin{aligned} 1-\beta &= 1-\mathbb{P}\left\{\text{type II error}\right\} = \mathbb{P}\left\{\text{accept } H_0 | \mu=\mu^\star\right\} \\ &= 1-\mathbb{P}\left\{14.3 \leq \bar{X}_n \leq 15.9 | \mu=\mu^\star\right\} \\ &= 1-\Phi(\frac{15.9-\mu^\star}{0.4}) + \Phi(\frac{14.3-\mu^\star}{0.4}), \end{aligned}$$

therefore power ($=1-\beta$) is a function of μ.

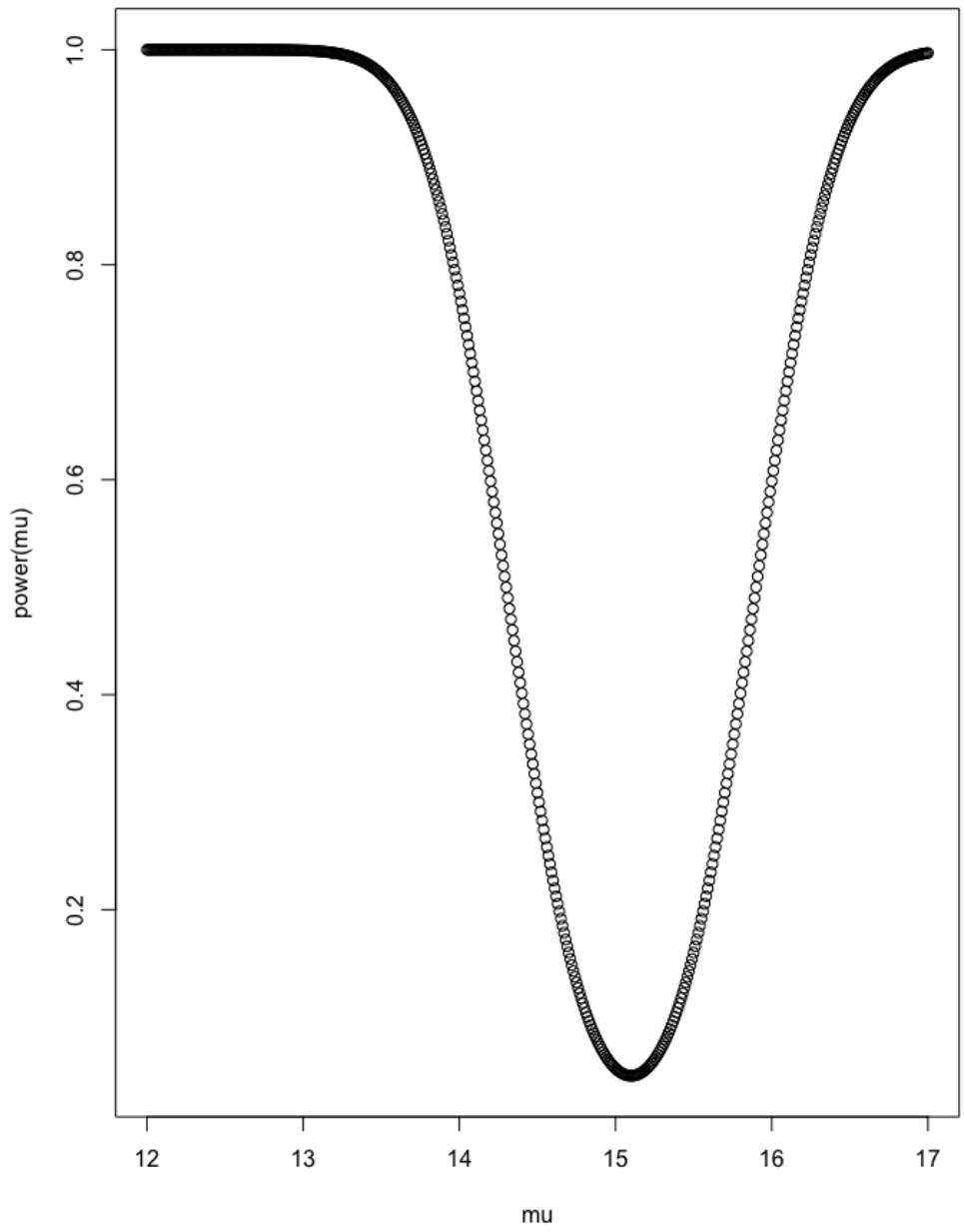

Figure: The power as a function of μ.

Hypothesis testing.

Example

Consider two coins, one black, one white, s.t.

$$\mathbb{P}_b\{\text{heads}\} = \mathbb{P}\{\text{heads}|\text{black coin thrown}\} = 0.5 \quad (1)$$

$$\mathbb{P}_w\{\text{heads}\} = \mathbb{P}\{\text{heads}|\text{white coin thrown}\} = 0.7. \quad (2)$$

One coin tossed $n = 10$ times, shows H heads.
Based on H, how would you decide which coin was used?

h	0	1	2	3	4	5
$\mathbb{P}_b\{H = h\}$	0.001	0.0098	0.0439	0.1172	0.2051	0.2461
$\mathbb{P}_w\{H = h\}$	0.0	0.0001	0.0014	0.009	0.0368	0.1029
h	6	7	8	9	10	
$\mathbb{P}_b\{H = h\}$	0.2051	0.1172	0.0439	0.0098	0.001	
$\mathbb{P}_w\{H = h\}$	0.2001	0.2668	0.2335	0.1211	0.0282	

Table: Probabilities of Binomial$(10, p)$ for $p = 0.5$ and $p = 0.7$.

Hypothesis testing.

Example 4. Based on H, how would you decide which coin was tossed?

h	0	1	2	3	4	5
$\mathbb{P}_b\{H=h\}$	0.001	0.0098	0.0439	0.1172	0.2051	0.2461
$\mathbb{P}_w\{H=h\}$	0.0	0.0001	0.0014	0.009	0.0368	0.1029
h	6	7	8	9	10	
$\mathbb{P}_b\{H=h\}$	0.2051	0.1172	0.0439	0.0098	0.001	
$\mathbb{P}_w\{H=h\}$	0.2001	0.2668	0.2335	0.1211	0.0282	

Table: Probabilities of Binomial$(10, p)$ for $p = 0.5$ and $p = 0.7$ (rounded to 3 decimals).

Suppose $H = 3$, then

$$\frac{\mathbb{P}_b\{H=3\}}{\mathbb{P}_w\{H=3\}} = \frac{0.1172}{0.009} \approx 13.0,$$

i.e. data suggest it is about 13 times more likely that black coin was thrown.

Hypothesis testing.

Example 4.

h	0	1	2	3	4	5
$\mathbb{P}_b\{H=h\}$	0.001	0.0098	0.0439	0.1172	0.2051	0.2461
$\mathbb{P}_w\{H=h\}$	0.0	0.0001	0.0014	0.009	0.0368	0.1029
h	6	7	8	9	10	
$\mathbb{P}_b\{H=h\}$	0.2051	0.1172	0.0439	0.0098	0.001	
$\mathbb{P}_w\{H=h\}$	0.2001	0.2668	0.2335	0.1211	0.0282	

Table: Probabilities of Binomial$(10, p)$ for $p = 0.5$ and $p = 0.7$ (rounded to 3 decimals).

On the other hand, if $H = 9$, then

$$\frac{\mathbb{P}_b\{H=9\}}{\mathbb{P}_w\{H=9\}} = \frac{0.0098}{0.1211} \approx 0.081,$$

i.e. data suggest it is about $1/0.081 \approx 12$ times more likely that white coin was thrown.

Hypothesis testing.

Example 4.

h	0	1	2	3	4	5
$\frac{\mathbb{P}_b\{H=h\}}{\mathbb{P}_w\{H=h\}}$	165.38	70.88	30.38	13.02	5.58	2.39
h	6	7	8	9	10	
$\frac{\mathbb{P}_b\{H=h\}}{\mathbb{P}_w\{H=h\}}$	1.02	0.44	0.19	0.08	0.03	

Table: Likelihood of Binomial$(10, 0.5)$ vs. Binomial$(10, 0.7)$ (rounded to 3 decimals).

Large values of so-called *likelihood ratio*

$$\frac{\mathbb{P}_b\{H = h\}}{\mathbb{P}_w\{H = h\}}$$

support

null hypothesis H_0: black coin was tossed (i.e. $p = 0.5$),

whereas small values support

alternative H_A: white coin was tossed (i.e. $p = 0.7$).

Hypothesis testing.

Example 4.

h	0	1	2	3	4	5
$\frac{\mathbb{P}_b\{H=h\}}{\mathbb{P}_w\{H=h\}}$	165.38	70.88	30.38	13.02	5.58	2.39
h	6	7	8	9	10	
$\frac{\mathbb{P}_b\{H=h\}}{\mathbb{P}_w\{H=h\}}$	1.02	0.44	0.19	0.08	0.03	

Table: Likelihood of Binomial$(10, 0.5)$ vs. Binomial$(10, 0.7)$ (rounded to 3 decimals).

The value $k = 6$ is critical in that

$$\begin{cases} \frac{\mathbb{P}_b\{H=h\}}{\mathbb{P}_w\{H=h\}} > 1 & \text{for } h \leq k \\ \frac{\mathbb{P}_b\{H=h\}}{\mathbb{P}_w\{H=h\}} < 1 & \text{for } h > k, \end{cases}$$

in other words, observing $h \leq 6$ (just) suggests that it is more likely that the black coin was tossed.

Hypothesis testing.

Example 4. Taking rejection region $C = \{7, 8, 9, 10\}$ yields a significance level

$$\alpha = \mathbb{P}\{\text{type I error}\} = \mathbb{P}_b\{H > 6\} = 0.18$$

with type II error probability

$$\beta = \mathbb{P}\{\text{type II error}\} = \mathbb{P}_w\{H \leq 6\} = 0.35.$$

If we are not willing to risk rejecting H_0 when, in fact, the black coin was tossed, with probability 0.18, but we are willing to risk this error with probability 0.01, i.e. $\alpha = 0.01$, we need to shrink the rejection region.
Setting rejection region to $C = \{9, 10\}$ yields

$$\alpha = \mathbb{P}_b\{H \geq 9\} = 0.01,$$

and type II error probability

$$\beta = \mathbb{P}_w\{H \leq 8\} = 0.85.$$

Example 4. More generally, for rejection region $C := \{k, \ldots, 10\}$ we obtain

$$\alpha = \mathbb{P}\{\text{type I error}\} = \mathbb{P}_b\{k, \ldots, 10\} = \sum_{i=k}^{10} \binom{10}{i} (\frac{1}{2})^{10},$$

and

$$\beta = \mathbb{P}\{\text{type II error}\} = \mathbb{P}_w\{1, \ldots, k-1\} = \sum_{i=1}^{k-1} \binom{10}{i} (0.7)^i (0.3)^{10-i}.$$

This calculation shows how reducing the type I error probability (by choosing a larger value for k) is at the cost of increasing the type II error probability, and vice versa.

Hypothesis testing.

Likelihood ratio test for simple hypotheses. The previous example is an instance of the so-called *likelihood ratio test*. In general, for any two simple hypotheses

$$H_0 \colon \theta = \theta_0 \qquad H_A \colon \theta = \theta_A$$

the *likelihood ratio (LR)* is defined to be

$$\Lambda := \frac{\mathrm{lik}(\theta_0)}{\mathrm{lik}(\theta_A)} = \frac{f(x_1, \ldots, x_n|\theta_0)}{f(x_1, \ldots, x_n|\theta_A)}.$$

The corresponding test with decision rule

$$\begin{cases} \text{reject } H_0 & \text{if } \Lambda < K \\ \text{accept } H_0 & \text{otherwise} \end{cases}$$

for some constant K is called a the *likelihood ratio test (LRT)*. As in the previous example, large values of Λ suggest that data support H_0, whereas small values suggest that data support H_A over H_0.

Hypothesis testing.

In general, even if we fix a significance level α, there can be a number of hypothesis tests with this level. Ideally, among all tests with level α we'd like to find one (there might be several) that minimizes the type II error probability, respectively maximizes power.
In the setting where both the null hypothesis and the alternative are simple, the next theorem shows that the optimal test (in the sense above) is the likelihood ratio test.

Hypothesis testing.

Lemma (Neyman-Pearson lemma)

Consider simple hypotheses H_0 and H_A and the corresponding likelihood ratio test with significance level α and power $1-\beta$. Then, any other test with significance level at most α has power less than or equal to $1-\beta$.

(without proof)

Example 5. Consider random sample $X_1, \ldots, X_n$ s.t. $X_1 \sim \mathcal{N}(\mu, \sigma^2)$ with known variance σ^2. The hypotheses are

$$H_0 \colon \mu = \mu_0 \qquad H_A \colon \mu = \mu_A$$

for some given $\mu_0 > \mu_A$. Likelihood ratio

$$\Lambda = \frac{\mathrm{lik}(\theta_0)}{\mathrm{lik}(\theta_A)} = \frac{e^{-\frac{1}{2}\sum_{i=1}^n (X_i - \mu_0)^2/\sigma^2}}{e^{-\frac{1}{2}\sum_{i=1}^n (X_i - \mu_A)^2/\sigma^2}} = e^{-\frac{1}{2\sigma^2}\sum_{i=1}^n [(X_i-\mu_A)^2 - (X_i-\mu_0)^2]}.$$

Here and in general, the likelihood ratio is a complicated statistic. However, one can often find a simpler statistic which determines the LR.

Example 5. Since

$$\sum_{i=1}^{n}[(X_i-\mu_A)^2-(X_i-\mu_0)^2] = -2\mu_0 n\bar{X}_n + n\mu_0^2 + 2\mu_A n\bar{X}_n - n\mu_A^2$$
$$= 2n\bar{X}_n(\mu_0-\mu_A) + n\mu_0^2 - n\mu_A^2,$$

the likelihood ratio is small for large values of $\bar{X}_n$, i.e. LRT rejects H_0 if

$$\Lambda < K \quad \text{or, equivalently, if} \quad \bar{X}_n > K'$$

for some values of K, K'.
In other words, instead of the LR we might as well use the sample mean $\bar{X}_n$ to construct the decision rule (replacing the constant K by K'.). This allows us to work out K, since we know the sampling distribution of $\bar{X}_n$.

Hypothesis testing.

To proceed, let us fix a significance level, say $\alpha = 0.01$. In order for the LRT to have significance level α, we have to solve

$$0.01 = \alpha = \mathbb{P}\{\text{reject } H_0 | H_0\} = \mathbb{P}\left\{\bar{X}_n > K' | H_0\right\}$$
$$= \mathbb{P}\left\{\frac{\bar{X}_n - \mu_0}{\sigma/\sqrt{n}} > \frac{K' - \mu_0}{\sigma/\sqrt{n}} | H_0\right\} = 1 - \Phi(\frac{K' - \mu_0}{\sigma/\sqrt{n}})$$

for K', i.e. $K' = \frac{\sigma}{\sqrt{n}}\Phi^{-1}(0.99) + \mu_0$, where $\Phi^{-1}(x)$ denotes the quantile function of the standard normal distribution.[1] For the power, we find

$$1 - \beta = \mathbb{P}\{\text{reject } H_0 | \mu = \mu_A\} = \mathbb{P}\left\{\bar{X}_n > K' | \mu = \mu_A\right\}$$
$$= 1 - \Phi(\frac{K' - \mu_A}{\sigma/\sqrt{n}})$$

According to the Neyman-Pearson lemma, there is no other test with significance level $\leq \alpha = 0.01$ that has a power greater than $1 - \Phi(\frac{K' - \mu_A}{\sigma/\sqrt{n}})$.

[1] The quantile function of the standard normal distribution is sometimes called the *probit* function. The corresponding command in R is qnorm.