

MS-A0503 First course in probability and statistics

6A Hypothesis testing

Jukka Kohonen

Department of mathematics and systems analysis
Aalto SCI

Academic year 2022–2023
Period III

Contents

Introduction to hypothesis testing

Examples and models

Variants

Decisions and errors

Further topics

Paul the Octopus

By choosing from two food boxes (with national flags), Paul predicted the winner of football matches. In 2008, correct 4/6 times. In 2010, correct 7/7 times.



| Opponent | Tournament | Stage | Date | Prediction | Result | Outcome |
|-----------|----------------|--------------------|--------------|----------------------------|--------|-----------|
| Poland | Euro 2008 | group stage | 8 June 2008 | Germany | 2–0 | Correct |
| Croatia | Euro 2008 | group stage | 12 June 2008 | Germany ^{[3][20]} | 1–2 | Incorrect |
| Austria | Euro 2008 | group stage | 16 June 2008 | Germany | 1–0 | Correct |
| Portugal | Euro 2008 | quarter-finals | 19 June 2008 | Germany | 3–2 | Correct |
| Turkey | Euro 2008 | semi-finals | 25 June 2008 | Germany | 3–2 | Correct |
| Spain | Euro 2008 | final | 29 June 2008 | Germany ^[3] | 0–1 | Incorrect |
| Australia | World Cup 2010 | group stage | 13 June 2010 | Germany ^[31] | 4–0 | Correct |
| Serbia | World Cup 2010 | group stage | 18 June 2010 | Serbia ^[31] | 0–1 | Correct |
| Ghana | World Cup 2010 | group stage | 23 June 2010 | Germany ^[31] | 1–0 | Correct |
| England | World Cup 2010 | round of 16 | 27 June 2010 | Germany ^[32] | 4–1 | Correct |
| Argentina | World Cup 2010 | quarter-finals | 3 July 2010 | Germany ^[23] | 4–0 | Correct |
| Spain | World Cup 2010 | semi-finals | 7 July 2010 | Spain ^[33] | 0–1 | Correct |
| Uruguay | World Cup 2010 | 3rd place play-off | 10 July 2010 | Germany | 3–2 | Correct |

Is this something that might easily happen by chance? Or does it indicate a good prediction skill?

https://en.wikipedia.org/wiki/Paul_the_Octopus

Hypothesis testing, contrasted to posterior inference

On previous lectures, we learned how we can infer a full distribution for an unknown parameter θ , if we have two ingredients:

- prior $f(\theta)$ — which values of θ are probable in the first place
- likelihood $f(\vec{x} | \theta)$ — the stochastic model of how the data are generated, if θ has a particular value

What if we are not able to formulate any prior $f(\theta)$? Can we do any inference **only from the data and the likelihood function**?

We can still do **something**. We can consider a **particular value of θ** , and choose to **reject** it, if that θ makes the observed data seem “too unlikely”. [We'll make this more precise.]

This leads to the classical hypothesis testing, which is the topic of this lecture. (This is an alternative to Bayesian inference.)

Hypothesis testing — first idea (not good)

Suppose we know the general stochastic model: $X \sim \text{Bin}(1000, \theta)$ (one thousand coin tosses), but don't know the parameter θ . We are considering if $\theta = 0.5$ seems plausible — or if the data seems too surprising (unlikely) for this parameter value.

Example 1. Observe $x = 510$ heads. If $\theta = 0.5$ is true,

$$\mathbb{P}(X = 510 | \theta = 0.5) = \binom{1000}{510} 0.5^{510} 0.5^{490} \approx 2.1\%.$$

Is this surprising? Should we reject $\theta = 0.5$?

Example 2. Observe $x = 500$ heads. If $\theta = 0.5$ is true,

$$\mathbb{P}(X = 500 | \theta = 0.5) = \binom{1000}{500} 0.5^{500} 0.5^{500} \approx 2.5\%.$$

Is this surprising? Should we reject $\theta = 0.5$? **Probably not!**

Hypothesis testing — classical method

| Step | Example |
|---|--|
| Formulate a hypothesis H_0 about how data are generated. | $\vec{X} = 30$ coffee cups, each from $N(10, 3^2)$ |
| Formulate a test statistic $t = t(\vec{X})$, calculated from data | Sample mean $m(\vec{X})$ |
| Work out the distribution of t (if H_0 is true). | $m(\vec{X}) \sim N(\dots)$ |
| Reject H_0 if the <i>observed</i> value $t(\vec{x})$ is in the tails of the distribution; choose tails to have probability α | $\alpha = 0.05$ |

Idea: t in the 5% tails is *surprising* if H_0 is true. We reject H_0 in that case. The tails are called **critical region** (**rejection region**).

Even if H_0 is true, this procedure may cause H_0 to be rejected — but only $\alpha = 5\%$ of the time. This is called the **significance level** of the test. → Illustration on blackboard

Hypothesis testing — another view, with p -value

| Step | Example |
|--|--|
| Formulate a hypothesis H_0 about how data are generated. | $\vec{X} = 30$ coffee cups, each from $N(10, 3^2)$ |
| Formulate a test statistic $t = t(\vec{X})$, calculated from data | Sample mean $m(\vec{X})$ |
| Work out the distribution of t (if H_0 is true). | $m(\vec{X}) \sim N(\dots)$ |
| Calculate both tail probabilities corresponding to $t(\vec{x})$. This is the p-value | $p = 0.018$ |
| Reject H_0 if $p < \alpha$ | reject |

Here we **first** calculated a p -value, and **then** applied the significance level $\alpha = 0.05$. A p -value $0.018 < 0.05$ was considered “surprising enough” that H_0 should be rejected.

Contents

Introduction to hypothesis testing

Examples and models

Variants

Decisions and errors

Further topics

Coffee machine — Normal model

A coffee machine is meant to give 10.0 cl coffee in each cup, at least on average. We assume the coffee volumes are normally distributed, but we don't know the mean. To test the hypothesis ($\mu = 10.0$), 30 cups were taken and measured:

11.05 9.65 10.93 9.46 10.27 10.02 10.07 10.74 11.15 10.40 10.12
11.20 10.07 10.27 9.99 9.80 10.83 10.21 11.26 10.11 10.49 10.10
10.15 11.02 10.00 11.68 10.51 11.20 11.29 10.15

Is the machine correctly calibrated (on average)?

Sample mean $m(\vec{x}) = 10.473$, which differs from the intended $\mu_0 = 10.0$.

But since the data are random, it is quite expected that the sample mean is not exactly 10.0!

Is the observed difference **statistically significant**?

Coffee machine — Normal model

11.05 9.65 10.93 9.46 10.27 10.02 10.07 10.74 11.15 10.40 10.12 11.20
10.07 10.27 9.99 9.80 10.83 10.21 11.26 10.11 10.49 10.10 10.15 11.02
10.00 11.68 10.51 11.20 11.29 10.15

Sample mean $m(\vec{x}) = 10.473$, sample standard deviation $sd(\vec{x}) = 0.563$

$H_0: \mu = \mu_0 = 10.0$

$H_1: \mu \neq \mu_0 = 10.0$

Test statistic of the observed data:

$$t(\vec{x}) = \frac{m(\vec{x}) - \mu_0}{sd(\vec{x})/\sqrt{n}} = \frac{10.473 - 10.0}{0.563/\sqrt{30}} = 4.60$$

Because sample size $n = 30$ fairly large, we work as if $\sigma = sd(\vec{x}) = 0.563$ exactly (“known variance”). Then $t(\vec{X})$ has standard normal distribution. [We could be more exact and use t distribution.]

$$p \text{ value} \approx \mathbb{P}\left(|t(\vec{X})| \geq |t(\vec{x})| \mid H_0\right) \approx \mathbb{P}(|Z| \geq 4.60) \approx 4.2 \times 10^{-6}$$

Result: p -value very small. If H_0 were true, it would be very unlikely to obtain a sample mean so far (or further) from the hypothesized $\mu = 10.0$.

Hypothesis testing vs. confidence interval

Often, hypothesis testing at **significance level α** can be alternatively framed as the question:

If we calculate a **$1 - \alpha$ confidence interval** for the unknown parameter θ , **does the interval contain** the value θ_0 claimed by the null hypothesis?

If the interval contains θ_0 , then the data is compatible with the possibility that $\theta = \theta_0$, as claimed.

If the interval is fully below or fully above θ_0 , then the data speaks against the possibility that $\theta = \theta_0$.

(Possibly illustration on blackboard)

Coffee machine — testing vs. confidence interval

11.05 9.65 10.93 9.46 10.27 10.02 10.07 10.74 11.15 10.40 10.12 11.20
10.07 10.27 9.99 9.80 10.83 10.21 11.26 10.11 10.49 10.10 10.15 11.02
10.00 11.68 10.51 11.20 11.29 10.15

Sample mean $m(\vec{x}) = 10.473$, sample standard deviation $sd(\vec{x}) = 0.563$

$$H_0: \mu = \mu_0 = 10.0$$

$$H_1: \mu \neq \mu_0 = 10.0$$

Again, work as if $\sigma = sd(\vec{x}) = 0.563$ exactly (“known variance”).

Computing e.g. 99% confidence interval, we obtain

$$10.473 \pm 2.58 \cdot \frac{0.563}{\sqrt{30}} \approx 10.473 \pm 0.265,$$

so the interval is completely above 10.0. Thus we reject the null hypothesis (that $\mu = 10.0$) at 1% significance level.

Caveat: In some situations, there are subtle differences between hypothesis testing and confidence intervals, but in the most common situations, this connection is probably helpful for understanding.

Null hypothesis H_0

The starting point of a hypothesis test is the **null hypothesis H_0** , which generally indicates that nothing new or surprising is needed to explain the observations. Often this is of the form “parameter=value” (and the most common parameter is *mean*).

Example

H_0 : Paul's predictions are correct with probability $\theta = 0.5$

H_0 : Coffee machine gives $\mu = 10.0$ cl on average, as intended

H_0 : A proposed new medicine is no better than placebo

H_0 : A portfolio manager performs no better than market average

The **alternative hypothesis H_1** is usually the complement of the null hypothesis. So if H_0 says $\mu = 10$, then H_1 says $\mu \neq 10$. Note that such an alternative hypothesis does not claim any single value!

Test statistic and p -value

The “surprisingness” of an observed data $\vec{x} = (x_1, \dots, x_n)$ is measured by first calculating a **test statistic**,

$$t(\vec{x}) = t(x_1, \dots, x_n),$$

which *condenses* the n -dimensional data vector into one real number.

Then the **p -value** (related to the test statistic) is the *probability* that the test statistic would have the observed value $t(\vec{x})$, **or something even further away** from the expected value.

The probability and the expected value are calculated by assuming that the H_0 is *true*. Some typical interpretations

| p-value | Interpretation |
|----------------|--------------------------------------|
| > 0.10 | Data quite compatible with H_0 |
| ≈ 0.05 | Data suggests against H_0 |
| < 0.01 | Data suggests strongly against H_0 |

Some more examples

Example (Coin tossing — Discrete data)

A coin that was claimed to be fair, was tossed 50 times, with 42 heads.

H_0 : Heads probability $\theta = 1/2$

H_1 : Heads probability $\theta \neq 1/2$

Example (Noisy observation — Little data)

Star brightness measurements claimed to be normal, with $\mu = 5$ and $\sigma = 3$. Measured once, with result $x_1 = 9.8$.

H_0 : $\mu = 5$

H_1 : $\mu \neq 5$

Example (Quality control — Composite hypothesis)

Shopkeeper claims that *at most* 5% of their tomatoes are bad. 50 tomatoes were tested, 7 were bad.

H_0 : Proportion of bad $\theta \leq 0.05$

H_1 : Proportion of bad $\theta > 0.05$

This is an example where H_0 is *composite* (allows many values).

Example. Coin tossing

Coin claimed to be fair, results 42 heads on 50 tosses.

H_0 : Heads probability $\theta = 1/2$

H_1 : Heads probability $\theta \neq 1/2$

Test statistic = heads count: $t(x) = 42$

$T = t(X) =$ "heads count according to H_0 "

$$f(x) = \mathbb{P}(T = x | H_0) = \binom{50}{x} \left(\frac{1}{2}\right)^x \left(1 - \frac{1}{2}\right)^{50-x}$$

Test statistic has mean $t_0 = \mathbb{E}(T | H_0) = 25$.

$$\begin{aligned} \text{p-value} &= \mathbb{P}\left(|T - t_0| \geq |t(x) - t_0| \mid H_0\right) \\ &= \mathbb{P}\left(|T - 25| \geq 17 \mid H_0\right) \\ &= \sum_{x=0}^8 f(x) + \sum_{x=42}^{50} f(x) \approx 1.2 \times 10^{-6}. \end{aligned}$$

Data is strongly against H_0 .

Example. Noisy observation

Star brightness measurements claimed to be normal $\mu = 5$ and $\sigma = 3$. Single observation: $x_1 = 9.8$.

H_0 : Mean $\mu = 5$

H_1 : Mean $\mu \neq 5$

Test statistic = normalized difference from the hypothesized mean:

$$z(\vec{x}) = \frac{x_1 - 5}{3} = 1.6$$

$$\text{p-value} = \mathbb{P}(|Z| \geq 1.6 \mid H_0) = 2\mathbb{P}(Z \geq 1.6 \mid H_0) \approx 11\%,$$

Observation compatible with regular random chance.

Observation does not lead to rejection of H_0

Contents

Introduction to hypothesis testing

Examples and models

Variants

Decisions and errors

Further topics

Variant: Testing for μ , large non-normal data

Suppose the data source generates independent, identically distributed numbers X_1, X_2, \dots, X_n from **some** distribution with unknown mean μ . We study whether the mean could be μ_0 .

$$H_0: \mu = \mu_0$$

$$H_1: \mu \neq \mu_0$$

Distribution unknown \implies impossible to test?

No; if sample is big, and independent, then CLT says the *sample mean* is normal, even if the individual observations are not.

Test statistic just like in the normal model:

$$t(\vec{x}) = \frac{m(\vec{x}) - \mu_0}{\text{sd}(\vec{x})/\sqrt{n}}.$$

Variant: Unknown variance

Often, the standard deviation σ **of the data source** is not known, but is estimated by the *sample* standard deviation $\text{sd}(\vec{x})$.

If the sample is large (e.g. $n > 30$), the estimate is decent, but ...

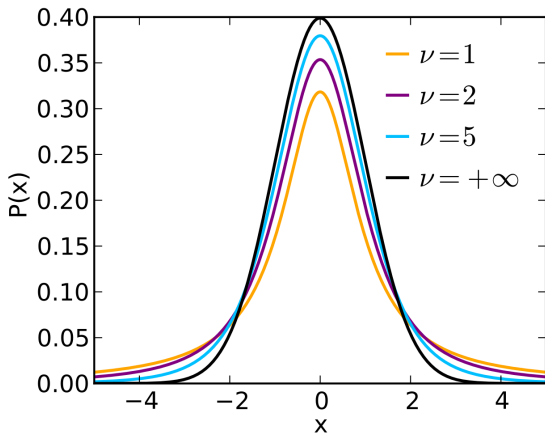
For small samples, we must note that the test statistic

$$t(\vec{X}) = \frac{m(\vec{X}) - \mu_0}{\text{sd}(\vec{X}) \sqrt{n}}$$

is the quotient of two random variables, and there is no reason to believe its distribution would be normal. It is not!

The real distribution of $t(\vec{X})$ is the **Student's t-distribution** with parameter $n - 1$. The parameter is called “degrees of freedom”. All is still fine — you simply do all computations with this t-distribution instead of the normal distribution. Again, you can use tables, or a computer. In R, `pt` is the CDF, and `qt` is the quantile function. (Compare to `pnorm` and `qnorm`.)

Student's t-distribution



Picture credit: Skbkekas, CC BY 3.0,

<https://commons.wikimedia.org/w/index.php?curid=9546828>

Interlude: Computing with distributions in R

| distribution | density | CDF | quantile function | generate random |
|--------------|---------|-------|-------------------|-----------------|
| uniform | dunif | pnif | qunif | runif |
| beta | dbeta | pbeta | qbeta | rbeta |
| normal | dnorm | pnorm | qnorm | rnorm |
| Student | dt | pt | qt | rt |
| exponential | dexp | pexp | qexp | rexp |
| ... | d... | p... | q... | r... |

Compare the 0.975-quantiles of standard normal, and Student with $n = 50$ and $n = 10$.

```
> qnorm(.975)
```

```
[1] 1.959964
```

```
> qt(.975, 49)
```

```
[1] 2.009575
```

=> Slightly wider confidence intervals.

```
> qt(.975, 9)
```

```
[1] 2.262157
```

=> Clearly wider confidence intervals.

Interlude: Computing with distributions in Matlab/Octave

| distribution | density | CDF | quantile function | generate random |
|--------------|---------|---------|-------------------|-----------------|
| uniform | unifpdf | unifcdf | unifinv | unifrnd |
| beta | betapdf | betacdf | betainv | betarnd |
| normal | normpdf | normcdf | norminv | normrnd |
| Student | tpdf | tcdf | tinv | trnd |
| exponential | exppdf | expcdf | expinv | exprnd |
| ... | ...pdf | ...cdf | ...inv | ...rnd |

Compare the 0.975-quantiles of standard normal, and Student with $n = 50$ and $n = 10$.

```
>> norminv(.975)
ans =
    1.959963984540054
>> tinv(.975, 49)
ans =
    2.009575237129235
>> tinv(.975, 9)
ans =
    2.262157162798204
```

Variant: Composite hypothesis

Shopkeeper claims that *at most* 5% of their tomatoes are bad.

50 tomatoes were tested, 7 were bad.

H_0 : Proportion of bad $\theta \leq 0.05$

H_1 : Proportion of bad $\theta > 0.05$

This is an example where H_0 is **composite** (allows many values).

Test statistic: Count of bads: $t(\vec{x}) = 7$

If the real proportion is θ (in the data source), then

$$\mathbb{P}_\theta(T = t) = f_\theta(t) = \binom{50}{t} \theta^t (1 - \theta)^{50-t}$$

Because H_0 claims proportion is *small*, we apply a **one-sided test**: only *high* values *above* claimed mean are significant. We would like to find

$$\mathbb{P}_\theta\left(T - \mathbb{E}_\theta(T) \geq t(\vec{x}) - \mathbb{E}_\theta(T)\right) = \mathbb{P}_\theta(T \geq t(\vec{x})) = \sum_{t=7}^{50} f_\theta(t).$$

Trouble: the probability depends on θ . So let us choose the the *highest* possible p -value, from any θ that H_0 allows:

$$p\text{-value} = \max_{\theta \leq 0.05} \mathbb{P}_\theta(T \geq t(\vec{x})) = \mathbb{P}_{0.05}(T \geq t(\vec{x})) = \sum_{t=7}^{50} f_{0.05}(t) \approx 1.2\%$$

Contents

Introduction to hypothesis testing

Examples and models

Variants

Decisions and errors

Further topics

Accepting or rejecting

You could compute a p -value and just report it, refraining of making further decisions like “accept” or “reject”.

But often you *need* to make a **decision**. Based on the test, you either accept or reject H_0 . This may affect e.g. further studies (performed or not), taking a medicine for use, ...

To make a decision, you choose (either before or after computing p) a **significance level** α ($0 < \alpha < 1$).

- If $p\text{-value} \geq \alpha$, the null hypothesis is *accepted*
- If $p\text{-value} < \alpha$, the null hypothesis is *rejected*

Typical, conventional significance levels are $\alpha = 1\%$ and $\alpha = 5\%$.

(This is a very crude way of making decisions. More advanced methods would explicitly consider the *consequences* of the decisions \rightarrow decision theory, but outside the scope of this course.)

Type I and II errors

Whichever decision we make (accept or reject), it may be correct or incorrect.

| | | Decision | |
|---------|-------------|----------------|----------------|
| | | H_0 accepted | H_0 rejected |
| Reality | H_0 true | Correct | Type I error |
| | H_0 false | Type II error | Correct |

If rejection of H_0 is considered *discovering* an interesting phenomenon (deviation from the null hypothesis), then

- type I error is a *false positive* (false discovery)
- type II error is a *false negative* (failure to discover)

In statistical inference, it is not possible to avoid both errors completely. But by probability calculus, we may try to calculate the *probabilities* of making type I and II errors.

Probabilities of the errors

$p(\vec{x})$ = the p-value computed from data \vec{x}

$p(\vec{X})$ = random variable: what p-values *can* be obtained (when \vec{X} follows a distribution)

If H_0 is true, then the probability of rejecting it (Type I error) is

$$\mathbb{P}(H_0 \text{ rejected} \mid H_0) = \mathbb{P}(p(\vec{X}) < \alpha \mid H_0) \approx \alpha$$

If H_0 is false, then the probability of accepting it (Type II error) is

$$\mathbb{P}(H_0 \text{ accepted} \mid H_1) = \mathbb{P}(p(\vec{X}) \geq \alpha \mid H_1)$$

By changing α , we can change both probabilities ... with a tradeoff

| α | Type I error rate | Type II error rate |
|----------|-------------------|--------------------|
| Small | Small | Large |
| Large | Large | Small |

Two caricatures

Eve Eager

- Applies significance level
 $\alpha = 5\%$
- Is eager to reject null hypotheses, so makes many discoveries
- Has approx 5% rate of Type I errors (rejecting a true null hypothesis)
- Has lower type II rate than Cathy

Cathy Cautious

- Applies significance level
 $\alpha = 1\%$
- Is cautious of rejecting a null hypothesis, so makes fewer discoveries
- Has approx 1% rate of type I errors (rejecting a true null hypothesis)
- Has higher type II error rate than Ann (failure to make a discovery)

Example. Coin tossing

A coin is tossed 10 times and $\vec{x} = (0, 0, 1, 0, 0, 0, 0, 0, 0, 0)$ is observed.
Test the fairness at significance 5%.

H_0 : Heads probability $\theta = 0.5$,

H_1 : Heads probability $\theta \neq 0.5$.

Test statistic: $t(\vec{x}) = \# \text{ heads}$

Stochastic model of the test statistic: $T = t(\vec{X})$

$$f_{H_0}(t) = \mathbb{P}(T = t | H_0) = \binom{10}{t} \left(\frac{1}{2}\right)^{10}$$

From this observed data \vec{x} , we compute

$$p(\vec{x}) = \mathbb{P}\left(|t(\vec{X}) - 5| \geq 4 \mid H_0\right) = \sum_{t=0}^1 f_{H_0}(t) + \sum_{t=9}^{10} f_{H_0}(t) \approx 2.1\%.$$

Decision: Null hypothesis rejected at 5% level.

But what do we know about the error probabilities?

Coin tossing, type I error rate (reject when H_0 TRUE)

Possible p-values, as a function of the test statistic $t(\vec{x}) = \# \text{heads}$:

| | | | | | | | | | | | |
|------------------|-----|-----|------|------|------|------|------|------|------|-----|-----|
| # heads | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| $f_{H_0}(t)$ [%] | 0.1 | 1.0 | 4.4 | 11.7 | 20.5 | 24.6 | 20.5 | 11.7 | 4.4 | 1.0 | 0.1 |
| p-value [%] | 0.2 | 2.1 | 10.9 | 34.4 | 75.4 | 100 | 75.4 | 34.4 | 10.9 | 2.1 | 0.2 |

At 5% level, we **reject** the null at the **critical region** $\{0, 1, 9, 10\}$.

If H_0 is true, we land there with probability

$$\mathbb{P}(t(\vec{X}) \in \{0, 1, 9, 10\} \mid H_0) = \sum_{t=0}^1 f_{H_0}(t) + \sum_{t=9}^{10} f_{H_0}(t) \approx 2.1\%.$$

So the type I error rate is $2.1\% \leq 5\%$.

It is not exactly 5% because in the discrete distribution of the test statistic, we do not have a point where the tail probabilities would be exactly 5%. Values 2 and 8 are in the acceptance region because their p-values are $> 5\%$.

Coin tossing, type II error rate (accept when H_0 FALSE)

Possible p-values, as a function of the test statistic:

| | | | | | | | | | | | |
|------------------|-----|-----|------|------|------|------|------|------|------|-----|-----|
| # heads | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| $f_{H_0}(t)$ [%] | 0.1 | 1.0 | 4.4 | 11.7 | 20.5 | 24.6 | 20.5 | 11.7 | 4.4 | 1.0 | 0.1 |
| $f_{H_1}(t)$ [%] | ? | ? | ? | ? | ? | ? | ? | ? | ? | ? | ? |
| p-value [%] | 0.2 | 2.1 | 10.9 | 34.4 | 75.4 | 100 | 75.4 | 34.4 | 10.9 | 2.1 | 0.2 |

At 5% level, we **accept** the null in the **complement of the critical region**, that is $\{2, 3, \dots, 7, 8\}$.

If H_1 is true, how probably do we land there (\Rightarrow type II error)?

This is more difficult to calculate, because it depends on the true value of θ , and H_1 allows many values.

For example, if $\theta = 0.5001$, we have

$$\begin{aligned}\mathbb{P}(t(\vec{X}) \in \{2, 3, \dots, 8\} \mid \theta = 0.5001) &\approx \mathbb{P}(t(\vec{X}) \in \{2, 3, \dots, 8\} \mid H_0) \\ &= \sum_{t=2}^8 f_{H_0}(t) \approx 97.9\%,\end{aligned}$$

so we have a huge type II error rate.

Type II error rate, if single alternative known

A coin tossed 10 times and $\vec{x} = (0, 0, 1, 0, 0, 0, 0, 0, 0, 0)$ is observed.

Extra assumption: We know that either $\theta = 0.5$ or $\theta = 0.9$. Test the fairness hypothesis at $\alpha = 0.05$.

H_0 : Heads probability $\theta = 0.5$,

H_1 : Heads probability $\theta = 0.9$.

Our computations are as before (same H_0 , same test statistic, same decisions). Now if H_1 is true, then the test statistic has distribution

$$f_{H_1}(t) = \mathbb{P}(T = t | H_1) = \binom{10}{t} 0.9^t (1 - 0.9)^{10-t}$$

and the type II error rate is

$$\mathbb{P}(t(\vec{X}) \in \{2, 3, \dots, 8\} | H_1) = \sum_{t=2}^8 f_{H_1}(t) \approx 26\%.$$

Contents

Introduction to hypothesis testing

Examples and models

Variants

Decisions and errors

Further topics

Further topics in hypothesis testing

The previous statistical tests concerned hypotheses *about the mean*, for example $p = 1/2$ or $\mu = 10.0$, and were based on *strong simplifying assumptions*, for example “data are normal” or “lots of data, so test statistic is normal”.

Classical statistics offers more tests for advanced questions, e.g.

- Hypotheses of **other parameters**. E.g. is the **standard deviation** of our star measurements $\sigma = 3$ or not?
→ χ^2 test etc.
- **Weaker assumptions**. E.g. data not normal and sample small, so sample mean not normal.
→ distribution-specific tests; or nonparametric tests
- Tests for **distribution shape**. E.g. we would like to test **whether** the data are normal. → more tests . . .

Further topics in hypothesis testing

For many specific yes/no questions about the unknown distribution (that generates the data), one can still apply the **same generic framework** of hypothesis testing:

1. Formulate a hypothesis H_0 about how the data are generated.
2. Formulate a test statistic $t(\vec{X})$ and work out its distribution, if H_0 is true.
3. Study how well the observed $t(\vec{x})$ fits into that distribution (is it in the tails or not).

Details of the test statistics and their distributions are different in each case.

More about such advanced tests e.g. on [MS-C1620 Statistical inference](#).

Last lecture on Friday, Feb 17. We will try to wrap up what we have learned during the course, see how it fits together, and perhaps fill in some gaps.

For the last lecture, you are encouraged to **bring your questions** about any topics related to the course. You can also send such questions in advance by e-mail.

Course exam on Wednesday, Feb 22.

- Time and place: See SISU
- Details: See course page → Announcements