Two-Sample Tests

Probability and Statistics for Data Science

Carlos Fernandez-Granda

These slides are based on the book Probability and Statistics for Data Science by Carlos Fernandez-Granda, available for purchase here. A free preprint, videos, code, slides and solutions to exercises are available at https://www.ps4ds.net

Hypothesis testing

1. Choose a conjecture
2. Choose null hypothesis
3. Choose test statistic
4. Decide significance level α
5. Gather data and compute test statistic
6. Compute p value
7. Reject the null hypothesis if p value $\leq \alpha$

Conjecture

Antetokounmpo's free throw percentage is different at home and away

Null hypothesis

Antetokounmpo's free throw percentage is the same at home and away

Two-sample test: Data are separated in two groups

Null hypothesis: Groups are from same distribution

Alternative hypothesis: Groups are from different distributions

Test statistic

Large value should be evidence against null hypothesis

$$t_{\text{data}} := \frac{\text{Made at home}}{\text{Attempted at home}} - \frac{\text{Made away}}{\text{Attempted away}}$$
$$= \frac{34}{44} - \frac{22}{41} = 0.236$$

Evidence against null hypothesis?

P value

Probability of observing larger or equal test statistic under null hypothesis

Two-sample z test

Data: $\tilde{x}_1, \ldots, \tilde{x}_n$

Two groups: A and B

One-tailed test statistic

$$\tilde{t}_{\text{1-tail}} = \frac{1}{n_A} \sum_{i \in \mathcal{A}} \tilde{x}_i - \frac{1}{n_B} \sum_{i \in \mathcal{B}} \tilde{x}_i$$

Null hypothesis: All data are i.i.d. Bernoulli with parameter θ_{null}

Binomial distribution

Binomial random variable $\tilde{a}$ with parameters n and θ

$\approx$ Gaussian with mean $n\,\theta$ and variance $n\,\theta\,(1-\theta)$

Two-sample z test

Null hypothesis: All data are i.i.d. Bernoulli with parameter θ_{null}

$$\tilde{t}_{\text{1-tail}} = \frac{1}{n_A} \sum_{i \in \mathcal{A}} \tilde{x}_i - \frac{1}{n_B} \sum_{i \in \mathcal{B}} \tilde{x}_i$$

Distribution of $\sum_{i \in \mathcal{A}} \tilde{x}_i$? Binomial with parameters n_A and θ_{null}

$\approx$ Gaussian with mean $n_A \theta_{\text{null}}$ and variance $n_A \theta_{\text{null}}(1 - \theta_{\text{null}})$

Gaussian random variable

If $\tilde{a}$ is a Gaussian random variable with mean μ and variance σ^2

$$\tilde{b} := \alpha \tilde{a} + \beta$$

is Gaussian with mean $\alpha\mu + \beta$ and variance $\alpha^2\sigma^2$

Two-sample z test

Distribution of $\sum_{i \in \mathcal{A}} \tilde{x}_i$?

$\approx$ Gaussian with mean $n_A \theta_{\text{null}}$ and variance $n_A \theta_{\text{null}}(1 - \theta_{\text{null}})$

Distribution of $\frac{1}{n_A} \sum_{i \in \mathcal{A}} \tilde{x}_i$?

$\approx$ Gaussian with mean θ_{null} and variance $\frac{\theta_{\text{null}}(1-\theta_{\text{null}})}{n_A}$

Distribution of $-\frac{1}{n_B} \sum_{i \in \mathcal{B}} \tilde{x}_i$?

$\approx$ Gaussian with mean $-\theta_{\text{null}}$ and variance $\frac{\theta_{\text{null}}(1-\theta_{\text{null}})}{n_B}$

Independent Gaussians $\tilde{a}$ and $\tilde{b}$

If $\tilde{a}_1$ and $\tilde{a}_2$ are independent Gaussian with means μ_1 and μ_2, and variances σ_1^2 and σ_2^2

$\tilde{s} = \tilde{a}_1 + \tilde{a}_2$ is Gaussian with mean $\mu_1 + \mu_2$ and variance $\sigma_1^2 + \sigma_2^2$

Two-sample z test

Null hypothesis: All data are i.i.d. Bernoulli with parameter θ_{null}

$$\tilde{t}_{\text{1-tail}} = \frac{1}{n_A} \sum_{i \in \mathcal{A}} \tilde{x}_i - \frac{1}{n_B} \sum_{i \in \mathcal{B}} \tilde{x}_i$$

Distribution of $\tilde{t}_{\text{1-tail}}$?

$\approx$ Gaussian with mean 0 and variance

$$\sigma^2_{\text{null}} := \theta_{\text{null}}(1 - \theta_{\text{null}}) \left(\frac{1}{n_A} + \frac{1}{n_B} \right)$$

In practice, $\theta_{\text{null}} := \frac{\text{Number of 1s}}{n}$

Antetokounmpo's free throws

$$
\begin{aligned}
t_{\text{data}} &:= \frac{\text{Made at home}}{\text{Attempted at home}} - \frac{\text{Made away}}{\text{Attempted away}} \\
&= \frac{34}{44} - \frac{22}{41} = 0.236
\end{aligned}
$$

Under null hypothesis $\approx$ Gaussian with mean 0 and variance

$$
\begin{aligned}
\sigma_{\text{null}} &:= \sqrt{\theta_{\text{null}}(1-\theta_{\text{null}})\left(\frac{1}{n_A}+\frac{1}{n_B}\right)} \\
&= \sqrt{\frac{56}{85}\left(1-\frac{56}{85}\right)\left(\frac{1}{44}+\frac{1}{41}\right)} = 0.103
\end{aligned}
$$

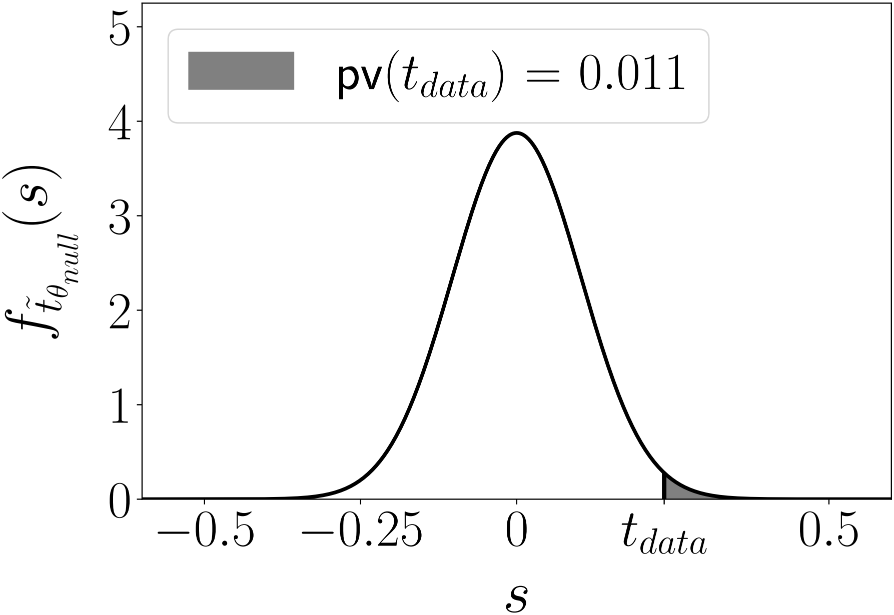
P value function

$\tilde{t}_{\text{1-tail}} \approx 0.103 \tilde{z}$

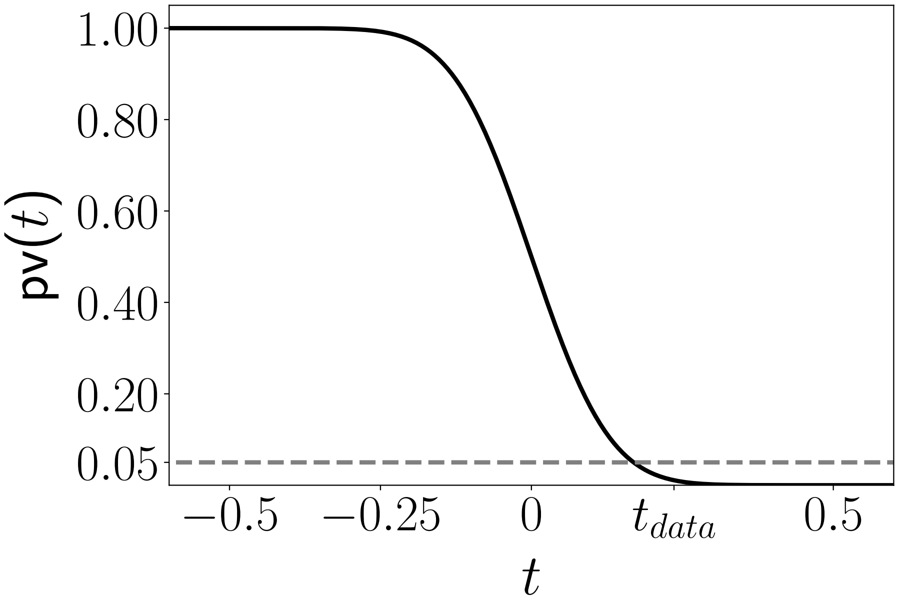
$\tilde{z}$ is standard Gaussian with mean 0 and variance 1

$$\begin{aligned} \text{pv}(t) &:= \mathrm{P}\left(\tilde{t}_{\text{1-tail}} \geq t\right) \\ &= \mathrm{P}\left(\tilde{z} \geq \frac{t}{0.103}\right) \end{aligned}$$

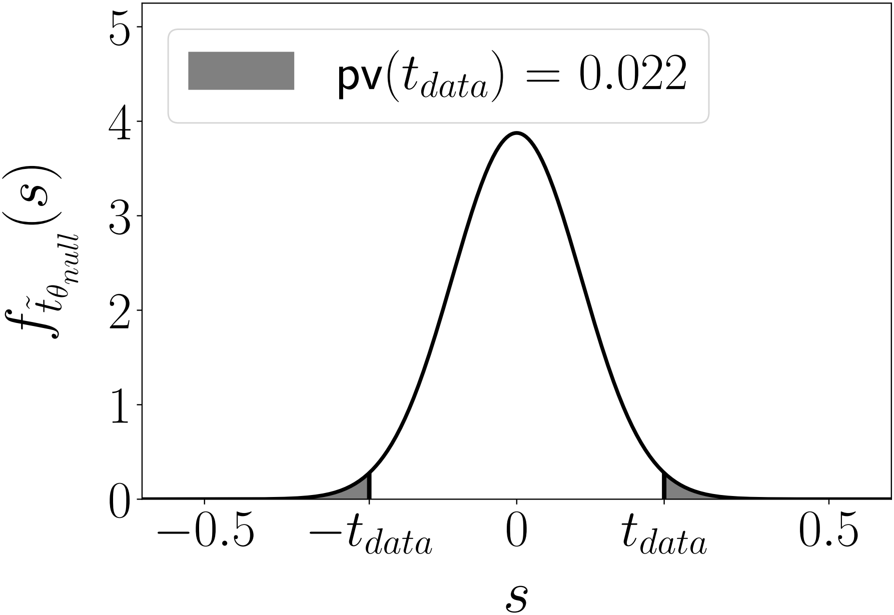
One-tailed test



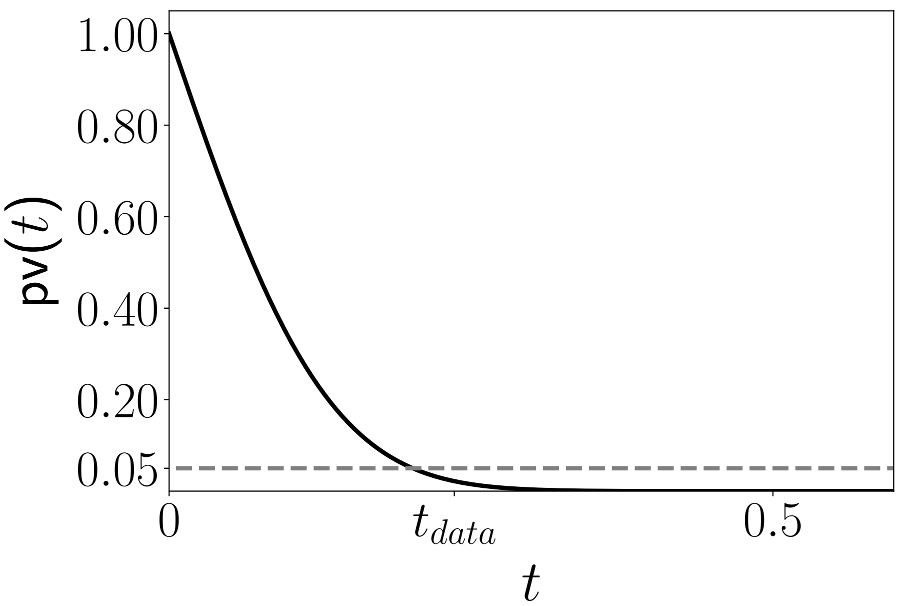
P value function



Two-tailed test



P value function



Statistical significance

We reject the null hypothesis when p value $\leq \alpha$

Guarantees that probability of a false positive $\leq \alpha$

Conclusion

$\alpha := 0.05 \geq 0.011$ (or 0.022)

We reject the null hypothesis!

Does this mean taunts cause worse percentage? No!

What have we learned

How to design a two-sample test