Multiple Testing

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 $\leq \alpha$

$\mathrm{P}\,(\text{False positive}) \leq \alpha$

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 $\leq \alpha$

Clutch

A player is clutch if they play better *when it matters*

Data: 3-point shooting during the 2014/2015 NBA season

Clutch time: 4th quarter of games decided by ≤ 10 points

Conjecture: Player shoots better in the clutch

Null hypothesis: Player shoots the same

Test statistic: 3s made in the clutch

Hypothesis test

Under null hypothesis, P(making a clutch 3) = season %

Distribution of test statistic $\tilde{t}_{\text{null}}$?

Binomial with parameters n and θ_{season}

P value

$$\begin{aligned} \text{pv}(t_{\text{data}}) &:= \mathrm{P}\left(\tilde{t}_{\text{null}} \geq t_{\text{data}}\right) \\ &= \sum_{i=t_{\text{data}}}^{n} \binom{n}{i} \theta_{\text{season}}^{i} \left(1-\theta_{\text{season}}\right)^{n-i} \end{aligned}$$

Significance level: $\alpha := 0.05$

	Season %	Clutch %	P value
Rob. Covington	38.2	73.3 (11/15)	0.006
Nikola Mirotic	34.1	62.5 (10/16)	0.019
Caron Butler	32.1	61.5 (8/13)	0.027
Mike Conley	39.2	60.9 (14/23)	0.029
Kirk Hinrich	31.7	52.4 (11/21)	0.039

Are you convinced?

2nd half of the season

	Season %	Clutch %	P value
Rob. Covington	38.2	31.8 (7/22)	0.796
Nikola Mirotic	34.1	37.5 (6/16)	0.478
Caron Butler	32.1	25.0 (2/8)	0.783
Mike Conley	39.2	50.0 (8/16)	0.262
Kirk Hinrich	31.7	37.5 (3/8)	0.491

What is going on?

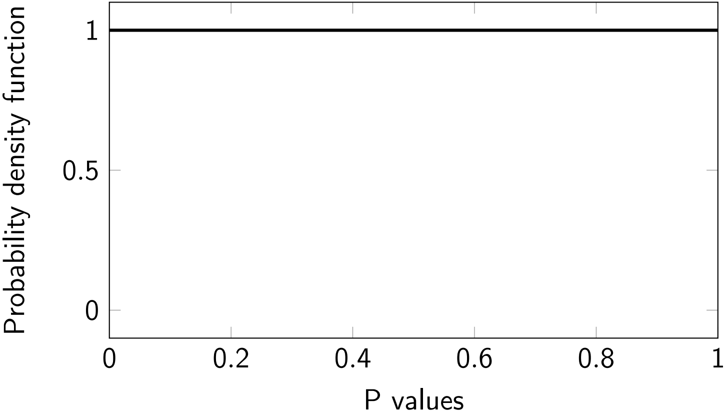
Probability that a single player overperforms by chance is low

But we are testing 146 players

Probability that a few of them overperform by chance is much higher!

P-value distribution

For continuous test statistics, distribution of p value under simple null hypothesis?

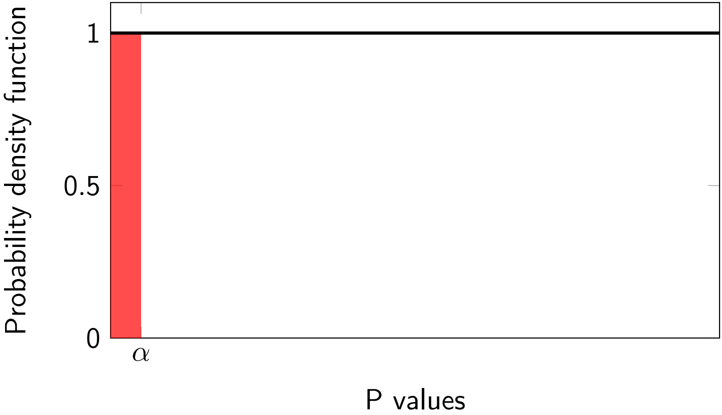


P-value distribution

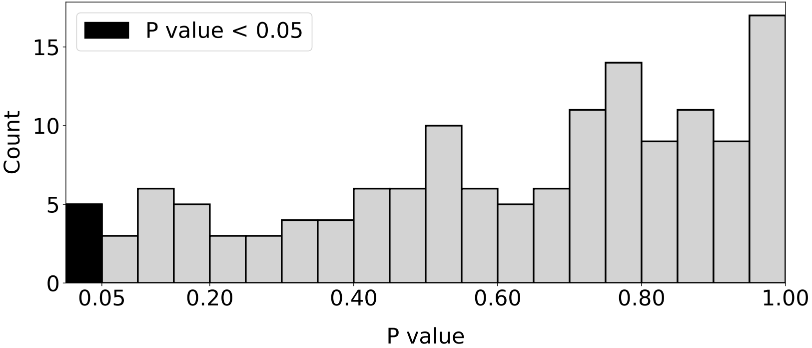
Probability of a single false positive? α

P-value distribution

Under null hypothesis, fraction of false positives among many tests? α!



P-value distribution for clutch example



Multiple testing

k independent hypothesis tests with significance level α

Probability of false positive in each test $= \alpha$

$$\begin{aligned} \mathrm{P}\left(\geq 1 \text{ false positive}\right) &= 1 - \mathrm{P}\left(\text{No false positives}\right) \\ &= 1 - (1-\alpha)^k \end{aligned}$$

For $\alpha := 0.05$ and $k := 100$, the probability is 0.99!

Solution? Decrease α

Challenge

How to set p-value threshold τ so that $\mathrm{P}\left(\text{False positive}\right) \leq \alpha$

$$\mathrm{P}\left(\text{False positive}\right) = \mathrm{P}\left(\cup_{i=1}^{k} \text{False positive in test } i\right)$$

Union bound

Events A_1, A_2, $\ldots A_k$

$$\mathrm{P}\left(\cup_{i=1}^{k} A_i\right) \leq \sum_{i=1}^{k} \mathrm{P}\left(A_i\right)$$

Bonferroni's correction

How to set p-value threshold τ so that $\mathrm{P}\left(\text{False positive}\right) \leq \alpha$

$$\begin{aligned}
\mathrm{P}\left(\text{False positive}\right) &= \mathrm{P}\left(\cup_{i=1}^{k} \text{False positive in test } i\right) \\
&\leq \sum_{i=1}^{k} \mathrm{P}\left(\text{False positive in test } i\right) \\
&\leq k\tau = \alpha
\end{aligned}$$

We reject null hypothesis if p value $\leq \tau := \alpha/k$

Guarantees $\mathrm{P}\left(\text{False positive}\right) \leq \alpha$

Clutch example

	Season %	Clutch %	P value
Rob. Covington	38.2	73.3 (11/15)	0.006
Nikola Mirotic	34.1	62.5 (10/16)	0.019
Caron Butler	32.1	61.5 (8/13)	0.027
Mike Conley	39.2	60.9 (14/23)	0.029
Kirk Hinrich	31.7	52.4 (11/21)	0.039

Bonferroni's threshold: $3.42 \cdot 10^{-4}$

Evaluating NBA players

Goal: Evaluate impact of a player on team performance

Statistic: Difference of mean point differential with/without player

$$t_{\text{data}} := m_{\text{with}} - m_{\text{without}}$$

2012-2018 NBA games

	Mean point diff.	Mins per game
Marcus Paige (CHA)	28.5	5.4
N. Mohammed (OKC)	18.5	4.0
Georges Niang (UTA)	17.1	3.7
L. James (CLE)	16.7	36.6
A. Goudelock (HOU)	16.5	6.4
B. Caboclo (TOR)	16.4	4.6
Roy Hibbert (DEN)	16.1	2.0
Brandon Knight (DET)	16.1	31.5
Michael Gbinije (DET)	15.8	3.4
DeMarre Carroll (BKN)	15.7	29.9

Hypothesis test

Null hypothesis: Player has no impact

Problem: No parametric model for test statistic under null hypothesis

Solution: Permutation test

Permutation test

Point differential x (n_1 games with / n_2 games without)

$$t_{\text{data}} = \text{mean}(x\,[1:n_1]) - \text{mean}(x\,[n_1+1:n_1+n_2])$$

We generate k permutations $v_1, \ldots, v_k \in \Pi_{x_{\text{data}}}$

$$T(v_i) = \text{mean}(v_i[1:n_1]) - \text{mean}(v_i[n_1+1:n_1+n_2])$$

$$\text{pv}(t_{\text{data}}) \approx \frac{\sum_{i=1}^{k} 1\,(T(v_i) \geq t_{\text{data}})}{k}$$

Are you convinced?

1,397 player/team pairs

	Mean point diff.	P value
Marcus Paige (CHA)	28.5	$2 \cdot 10^{-4}$
N. Mohammed (OKC)	18.5	$3 \cdot 10^{-3}$
Georges Niang (UTA)	17.1	$2 \cdot 10^{-4}$
L. James (CLE)	16.7	$< 10^{-7}$
A. Goudelock (HOU)	16.5	$3 \cdot 10^{-2}$
B. Caboclo (TOR)	16.4	$< 10^{-7}$
Roy Hibbert (DEN)	16.1	$3 \cdot 10^{-3}$
Brandon Knight (DET)	16.1	$2 \cdot 10^{-3}$
Michael Gbinije (DET)	15.8	$5 \cdot 10^{-3}$
DeMarre Carroll (BKN)	15.7	$2 \cdot 10^{-3}$

Bonferroni's correction

1,397 player/team pairs

If $\alpha := 0.05$, Bonferroni's threshold is $\alpha / k = 3.58 \cdot 10^{-5}$

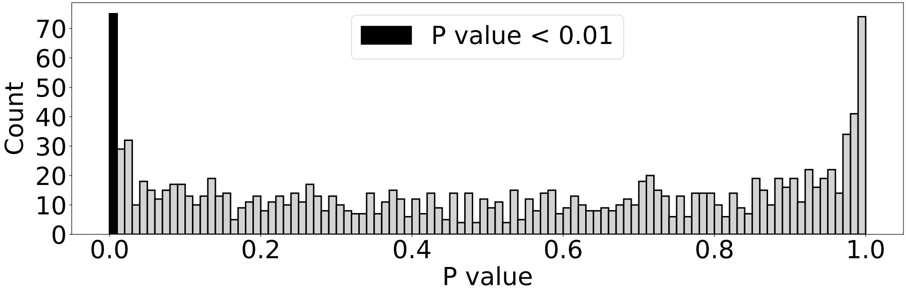
	Mean point diff.	P value
Marcus Paige (CHA)	28.5	$2 \cdot 10^{-4}$
N. Mohammed (OKC)	18.5	$3 \cdot 10^{-3}$
Georges Niang (UTA)	17.1	$2 \cdot 10^{-4}$
L. James (CLE)	16.7	$< 10^{-7}$
A. Goudelock (HOU)	16.5	$3 \cdot 10^{-2}$
B. Caboclo (TOR)	16.4	$< 10^{-7}$
Roy Hibbert (DEN)	16.1	$3 \cdot 10^{-3}$
Brandon Knight (DET)	16.1	$2 \cdot 10^{-3}$
Michael Gbinije (DET)	15.8	$5 \cdot 10^{-3}$
DeMarre Carroll (BKN)	15.7	$2 \cdot 10^{-3}$

Sorting by p values

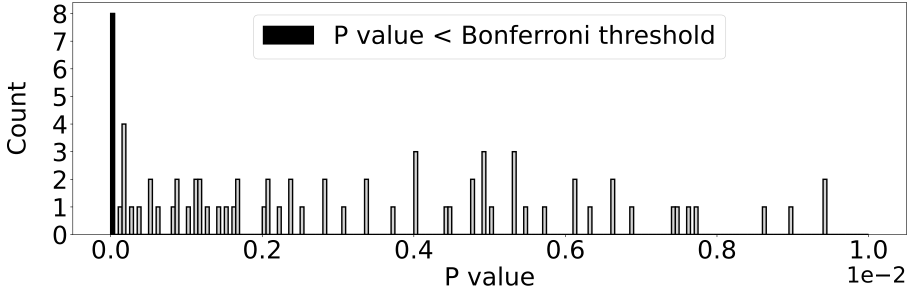
Bonferroni's threshold: $3.58 \cdot 10^{-5}$

	Mean point diff.	P value	Mins per game
L. James (CLE)	16.7	$< 10^{-7}$	36.6
B. Caboclo (TOR)	16.4	$< 10^{-7}$	4.6
N. Mirotic (CHI)	10.3	$3 \cdot 10^{-7}$	23.1
C. Anthony (NY)	8.1	$5 \cdot 10^{-7}$	36.3
Ricky Rubio (MIN)	7.6	$7 \cdot 10^{-7}$	31.4
James Jones (MIA)	8.2	$6 \cdot 10^{-6}$	7.8
Brandon Rush (GS)	6.7	$6 \cdot 10^{-6}$	12.6
Joel Embiid (PHI)	8.7	$2 \cdot 10^{-5}$	28.7
Kevin Durant (OKC)	6.9	$1 \cdot 10^{-4}$	37.3
Kevin Garnett (MIN)	9.2	$2 \cdot 10^{-4}$	15.3

P-value distribution



P-value distribution



False negative

Bonferroni's threshold: $3.58 \cdot 10^{-5}$

	Mean point diff.	P value	Mins per game
L. James (CLE)	16.7	$< 10^{-7}$	36.6
B. Caboclo (TOR)	16.4	$< 10^{-7}$	4.6
N. Mirotic (CHI)	10.3	$3 \cdot 10^{-7}$	23.1
C. Anthony (NY)	8.1	$5 \cdot 10^{-7}$	36.3
Ricky Rubio (MIN)	7.6	$7 \cdot 10^{-7}$	31.4
James Jones (MIA)	8.2	$6 \cdot 10^{-6}$	7.8
Brandon Rush (GS)	6.7	$6 \cdot 10^{-6}$	12.6
Joel Embiid (PHI)	8.7	$2 \cdot 10^{-5}$	28.7
Kevin Durant (OKC)	6.9	$1 \cdot 10^{-4}$	37.3
Kevin Garnett (MIN)	9.2	$2 \cdot 10^{-4}$	15.3

False negatives

Bonferroni's threshold: $3.58 \cdot 10^{-5}$

	Mean point diff.	P value	Mins per game
Marcus Paige (CHA)	28.5	$2 \cdot 10^{-4}$	5.4
Georges Niang (UTA)	17.1	$2 \cdot 10^{-4}$	3.7
Chris Paul (LAC)	6.8	$2 \cdot 10^{-4}$	33.6
Stephen Curry (GS)	8.2	$3 \cdot 10^{-4}$	34.6
Anthony Davis (NO)	5.1	$4 \cdot 10^{-4}$	34.8
Marc Gasol (MEM)	5.5	$5 \cdot 10^{-4}$	33.9
DeMarre Carroll (ATL)	10.1	$5 \cdot 10^{-4}$	31.5
Kawhi Leonard (SA)	4.7	$6 \cdot 10^{-4}$	31.6
Nikola Pekovic (MIN)	5.0	$8 \cdot 10^{-4}$	28.7
Klay Thompson (GS)	10.0	$9 \cdot 10^{-4}$	34.1

Tradeoff

Bonferroni's correction reduces false positives

But increases false negatives!

More sophisticated approaches order by p value and accept a certain fraction of false positives

What have we learned

Challenges arising from multiple testing

Bonferroni's correction

Tradeoff between false positives and false negatives

Wait a minute

	Mean point diff.	P value	Mins per game
L. James (CLE)	16.7	$< 10^{-7}$	36.6
B. Caboclo (TOR)	16.4	$< 10^{-7}$	4.6
N. Mirotic (CHI)	10.3	$3 \cdot 10^{-7}$	23.1
C. Anthony (NY)	8.1	$5 \cdot 10^{-7}$	36.3
Ricky Rubio (MIN)	7.6	$7 \cdot 10^{-7}$	31.4
James Jones (MIA)	8.2	$6 \cdot 10^{-6}$	7.8
Brandon Rush (GS)	6.7	$6 \cdot 10^{-6}$	12.6
Joel Embiid (PHI)	8.7	$2 \cdot 10^{-5}$	28.7

Played 24 games over 4 years (missing 200)

Were Raptors winning *because* Caboclo was playing?

Caboclo was playing because Raptors were winning