Hypothesis testing 2: p-value, GLRT test

Y. Polyanskiy, D. Shah, J. Tsitsiklis

6.S077

2018

Outline:

- Recap (basics, z-test, t-test
- p-value
- GLRT test
- **G-statistic**
- **Kolmogorov-Smirnov test**

Recall: Formal setting for HT

Definition

Statistical hypotheses:

- H : data $X_1, \ldots, X_n$ distributed according to $P \in \mathcal{C}_0$
- K : data $X_1, \ldots, X_n$ distributed according to $P \in \mathcal{C}_1$

where $\mathcal{C}_0, \mathcal{C}_1$ are COLLECTIONS OF DISTRIBUTIONS.

Remarks:

- **ALWAYS (!)** make sure you can formulate in the form above
- Most common special case: X_i are i.i.d. from P
- ... So will just write things like:

$$H : \mathbb{E}[X] = 0 \qquad \text{vs.} \qquad K : \mathbb{E}[X] \neq 0\,.$$

- Can reject H, but not "prove it"!

Design of tests

- **Before** seeing the data we announce test:

  Data $\boldsymbol{X} = (X_1, \ldots, X_n)$ *lands in set* $R_\alpha \Rightarrow$ *REJECT* *null.*

  with $P[\text{data} \in R_\alpha | H] \leq \alpha$ (false positive, significance)
- More exactly:

$$(*) \qquad P[\boldsymbol{X} \in R_\alpha] \leq \alpha \qquad \forall P \in \mathcal{C}_0$$

- How is R_α selected?
- Usually: by thresholding some statistic:

$$R_\alpha = \{T(\boldsymbol{X}) \geq t_\alpha\}\,.$$

- Statistic $T(\boldsymbol{X})$ is chosen with two goals in mind:
  - "pivotality": Under any null $P \in \mathcal{C}_0$ distribution of $T(\boldsymbol{X})$ is same
  - "consistency": Under any non-null $P \in \mathcal{C}_1$, $T(\boldsymbol{X})$ grows to ∞ with n
- Threshold t_α is selected to satisfy $(*)$ (approximately or **exactly**)
- We learned about two statistics with such properties: Z and T

Recap: z- and t-tests

Testing for mean

- Data $X_i \overset{iid}{\sim} P$, mean $\mathbb{E}[X] = \mu$
- null $H : \mu = 0$
- alt $K : \mu \neq 0$ (or $\mu > 0$)

- Good idea: compute sample mean $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i$
- z-test:
  - Applicable when $\mathrm{Var}[X|H] = \sigma_0^2$ (known, null-variance)
  - $\boxed{Z = \sqrt{\frac{n}{\sigma_0^2}}(\hat{\mu} - \mu_0)}$
  - Asymptotically normal: $n \gg 1$ have $Z \approx \mathcal{N}(0,1)$ under null!
  - REJECT null if $|Z| > 1.96$ for significance $\alpha = 0.05$
- t-test:
  - Applicable when variance is unknown (aka nuisance parameter)
  - $\boxed{T = \sqrt{\frac{n}{\widehat{\sigma^2}}}(\hat{\mu} - \mu_0)}$
  - Asymptotically normal: $n \gg 1$ have $T \approx \mathcal{N}(0,1)$ under null!
  - REJECT null if $|T| > 1.96$ for significance $\alpha = 0.05$

HT steps (again)

Hypothesis testing mindset:

1. Suppose in your experiment you will see data $\boldsymbol{X} = (X_1, \ldots, X_n)$
2. Formulate null hypothesis H: $X \sim P$ with $P \in \mathcal{C}_0$
3. Formulate alternative hypothesis $K : X \sim P$ with $P \in \mathcal{C}_1$
4. Choose statistic whose distribution under H is same for all $P \in \mathcal{C}_0$

$$T = \sqrt{n}\frac{\hat{\mu} - \mu_0}{\sqrt{\widehat{\sigma^2}}} \approx \mathcal{N}(0, 1)$$

5. Threshold test: If T "large", REJECT null H.
6. Threshold chosen s.t. $\mathbb{P}[\text{reject}|H] \leq \alpha$ for pre-specified α (typ. 0.05)
7. Only then see the data

Q: Do we really need 5-6? Why threshold at all?

Concept of p-value

- Consider test procedure:

$$(*) \qquad T(X) \geq t_\alpha \quad \Rightarrow \quad \text{reject null } H$$

- Value of T, aka "effect size", is more important than binary decision
- **Question:** "effect size" has units, can we convert it to universal scale?
- **Answer:** Yes! p-value is the answer.
- Can be computed if: 1) H is specified, 2) test is of the form $(*)$

Algorithm for computing p-value

- Got data $\boldsymbol{x} = (x_1, \ldots, x_n)$.
- Compute observed statistics $t_{obs} \triangleq T(\boldsymbol{x})$
- p-value $\triangleq P[T(\boldsymbol{X}) \geq t_{obs} | H]$

- Mnemonic: probability of observing same or more extreme data
- For $P[\cdot|H]$ to make sense, should have "pivotality" (for general case, wait a bit)

Illustration of p-value: $\mu = \mu_0$ vs $\mu > \mu_0$

Right-tailed test

Reject when $\{T(x) > t_\alpha\}$

p-value $\triangleq \mathbb{P}[T(X) \geq t_{obs} | H]$

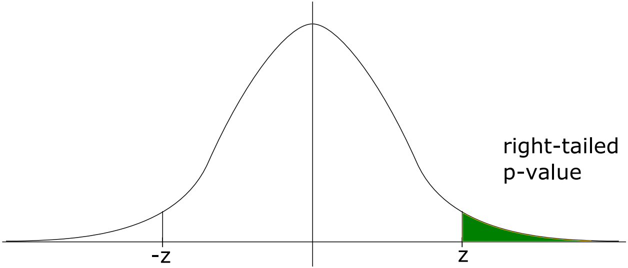


Illustration of p-value: $\mu = \mu_0$ vs $\mu \neq \mu_0$

Two-tailed test

Reject when $\{|T(x)| > t_{\frac{\alpha}{2}}\}$

p-value $\triangleq \mathbb{P}[|T(X)| \geq t_{obs}|H]$

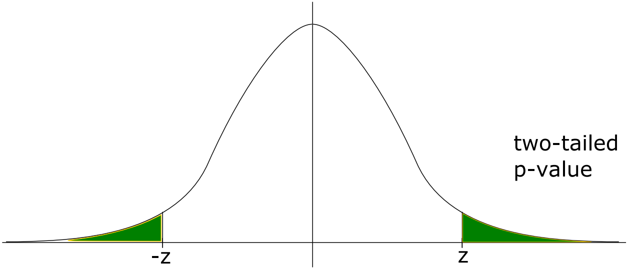
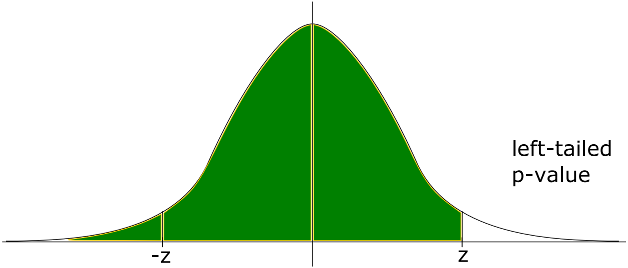


Illustration of p-value: $\mu = \mu_0$ vs $\mu < \mu_0$

<u>Left-tailed test</u>

Reject when $\{T(x) < t_\alpha\}$

p-value $\triangleq \mathbb{P}[T(X) \leq t_{obs} | H]$



Lady tasting tea

Back story

- M. Bristol claims to be able to tell whether tea or milk was poured into cup first
- Famous statistician R. Fisher is her colleague
- Proposes to test it

The experiment

- Design: 8 teacups are placed (4 tea first, 4 milk first)
- Data: X = tasting, Y = truth
  (e.g. $X = \texttt{TTTTMMMM}$, $Y = \texttt{TMTMTMTM}$)
- Test statistic: $T = \#$ of correct guesses
- Experiment: M. Bristol got $T = 8$. What is p-value?

Lady tasting tea

- Design: 8 teacups are placed (4 tea first, 4 milk first)
- Data: X = tasting, Y = truth
  (e.g. $X = \texttt{TTTTMMMM}$, $Y = \texttt{TMTMTMTM}$)
- Test statistic: $T = \#$ of correct guesses
- Experiment: M. Bristol got $T = 8$. What is p-value?

Hypothesis testing formulation

- Null hypothesis $H : X, Y$ are i.i.d. uniform on $\binom{8}{4} = 70$ strings
- Distribution of T under null:

T	Prob	T	Prob
0	$1/70$	6	$16/70$
2	$16/70$	8	$1/70$
4	$36/70$		

p-value $= 1/70 \approx 0.014$

Concept of p-value: MIT version

- Consider test procedure:

$$(*) \qquad T(\boldsymbol{X}) \geq t_\alpha \quad \Rightarrow \quad \text{reject null } H$$

- Value of T, aka "effect size", is more important than binary decision
- Then we have $t_{obs} \triangleq T(\boldsymbol{x})$ (observed value of T)

$$p\text{-value} \triangleq \mathbb{P}[T(\boldsymbol{X}) \geq t_{obs} | H]$$

- **Problem:** What is $\mathbb{P}[\cdot|H]$?
- **Solution:** Just replace with $\max_{P \in \mathcal{C}_0}$!

Algorithm for computing p-value

- Got data $\boldsymbol{x} = (x_1, \ldots, x_n)$.
- Compute observed statistics $t_{obs} \triangleq T(\boldsymbol{x})$
- p-value $\triangleq \max_{P \in \mathcal{C}_0} P[T(\boldsymbol{X}) \geq t_{obs}]$

Concept of p-value: MIT version (PhD's only)

- Consider test procedure:

$$(*) \qquad T(\boldsymbol{X}) \geq t_\alpha \quad \Rightarrow \quad \text{reject null } H$$

- Value of T, aka "effect size", is more important than binary decision
- Then we have $t_{obs} \triangleq T(\boldsymbol{x})$ (observed value of T)

$$\boxed{p\text{-value} \triangleq \max_{P \in \mathcal{C}_0} P[T(\boldsymbol{X}) \geq t_{obs}]}$$

- **Problem:** What if test is not of the form "$T(x) \geq t_\alpha$"?
- **Solution:** Let R_α be a family of tests s.t.

$$P[\boldsymbol{X} \in R_\alpha] \leq \alpha \quad \forall P \in \mathcal{C}_0$$

then

$$p\text{-value} \triangleq \inf\{\alpha : \boldsymbol{x} \in R_\alpha\}$$

REMEMBER: p-value is not a function of data **ONLY**.

- It depends on data and test.
- You cannot say this data is significant to reject null with $p = 0.001$.
- ... the test for which p is computed should be specified.
- In practice hard to decipher from actual papers.
- "Reproducible research" movement is to fix this.

Interpreting p-value

- Roughly: p-value $= P[\text{data (or more extrem)}|H]$
- Value $p = 0.05$ means false REJECT in 5% of experiments
- ... often used to decide on funding, continuing drug trials etc
- Rookie mistake: think p-value is $P[H|\text{data}]$
  Mass media: "null is true w.p. $< 5\%$"
- Fun calculation: If $p < 0.05$ is rejection threshold

  False-positive in 5 tests w.p. 22%
  False-positive in 10 tests w.p. 40%
  False-positive in 15 tests w.p. 53%
  False-positive in 20 tests w.p. 64%
- Recall: 10^6 articles per year in PubMed... so 50000 false positives
- ... and 2500 false-positive replications, 125 triple replications, 6 quad replications
- Sensational (false?) positives get blown up by the media

Data-snooping

Another rookie blunder

- Null: $\mu = \mu_0$
- See data, observed t-statistic $t_{obs} > 0$

  (i.e. sample-mean $> \mu_0$)
- Decide to report one-sided p-value.

  I.e. write paper "On testing $\mu = \mu_0$ vs. $\mu > \mu_0$"
- ERROR: cannot pick hypothesis after seeing data
- You report: "My p-value was calculated as"

$$p_{cheat} = \mathbb{P}[T(X) > t_{obs} | H]$$

  but in truth you computed

$$p_{true} = \mathbb{P}[T(X) > t_{obs} | H, T(X) > 0] \,.$$

- Under normal approximation $p_{true} \approx 2p_{cheat}$
- Example of data-snooping (beginner-level)
- Mid-level: do multiple trials, report one

  (Chicago Bears, coin tosses "$p = 2^{-14}$"?)
- Pro-level: run many tests, report one

Roadmap of tests

Tests we will learn:

- One-sample tests:
  1. for mean of population: $\mathbb{E}[X] = \mu_0$ vs $\mathbb{E}[X] \neq \mu_0$
  2. for other parameters: $\theta \in \Theta_0$ vs $\theta \notin \Theta_0$
  3. generalized likelihood-ratio test: $X \sim$ Uniform vs $X \sim$ not Uniform
  4. testing normality: $X \sim \mathcal{N}(0,1)$ vs $X \nsim \mathcal{N}(0,1)$

- Two-sample tests:
  1. Equality of means: $\mathbb{E}[X] = \mathbb{E}[Y]$ vs. $\mathbb{E}[X] \neq \mathbb{E}[Y]$
  2. Equality of distributions: $P_X = P_Y$ vs. $P_X \neq P_Y$
  3. Testing independence: $X \perp\!\!\!\perp Y$ vs $X \not\perp\!\!\!\perp Y$

Design of tests

- **Before** seeing the data we announce test:

  Data $\boldsymbol{X} = (X_1, \ldots, X_n)$ *lands in set* $R_\alpha \Rightarrow$ *REJECT null.*

- Usual choice of crit. region:

$$R_\alpha = \{T(\boldsymbol{X}) \geq t_\alpha\}.$$

- Statistic $T(\boldsymbol{X})$ is chosen with two goals in mind:
  - "pivotality": Under any null $P \in \mathcal{C}_0$ distribution of $T(\boldsymbol{X})$ is known
  - "consistency": Under any non-null $P \in \mathcal{C}_1$, $T(\boldsymbol{X})$ grows to ∞ with n
- How does one find such T????
- Art... (as in beautiful, cf. **exact** non-parametric tests)
- Some guidelines:
  - Use good $\hat{\theta}$
  - Shed nuisance scale parameters by Studentization
- How about cases other than $\theta \in H$ vs $\theta \in K$?

Generalized likelihood-ratio test

- How do we test for general hypotheses?
- MLE was our savior in estimation. Analog for HT?

The G-statistic

$$\begin{aligned} G &\triangleq -2\log\frac{P_0^*(x_1,\ldots,x_n)}{P_1^*(x_1,\ldots,x_n)} \\ P_0^*(x_1,\ldots,x_n) &= \max_{P\in\mathcal{C}_0} P(x_1,\ldots,x_n) \\ P_1^*(x_1,\ldots,x_n) &= \max_{P\in\mathcal{C}_0\cup\mathcal{C}_1} P(x_1,\ldots,x_n) \end{aligned}$$

- The GLRT: REJECT if $G > g_\alpha$
- Rationale: Large P_0/P_1 means H is more likely than K.
  Later: Neyman-Pearson Lemma
- Version with $\max_{P\in\mathcal{C}_1}$ is also useful
- Distribution of G under null? Let's find out ...

Testing for discrete distribution (goodness-of-fit)

HT problem

- X be r-valued: $[r] = \{1, \ldots, r\}$
- P_0 – a pmf on $[r]$, i.e. $P_0(1) + \cdots + P_0(r) = 1$
- Null $H : X \overset{iid}{\sim} P_0$
- Alt $K : X \overset{iid}{\sim} P$ with $P \neq P_0$

Derive G-statistic

- Let $\hat{P}(\cdot) = \frac{1}{n}\#\{j : x_j = \cdot\}$ empirical dist
- $P_0^*(x_1, \ldots, x_n) = \prod_{a=1}^r P_0(a)^{n\hat{P}(a)}$
- $P_1^*(x_1, \ldots, x_n) = \max_P \prod_{a=1}^r P(a)^{n\hat{P}(a)} = \prod_{a=1}^r \hat{P}(a)^{n\hat{P}(a)}$
- $G = 2n \sum_{a=1}^r \hat{P}(a) \log \frac{\hat{P}(a)}{P_0(a)}$
- $\ldots = 2nD(\hat{P} \| P_0)$ – distance-like measure of proximity (KL-divergence)
- **Strong MAGIC:** as $n \to \infty$ under null

$$G \approx \chi^2(r-1)$$ regardless of P_0!

Testing for discrete distribution (goodness-of-fit)

HT problem

- X be r-valued: $[r] = \{1, \ldots, r\}$
- P_0 – a pmf on $[r]$, i.e. $P_0(1) + \cdots + P_0(r) = 1$
- Null $H : X \overset{iid}{\sim} P_0$
- Alt $K : X \overset{iid}{\sim} P$ with $P \neq P_0$

- $G = 2n \sum_{a=1}^{r} \hat{P}(a) \log \frac{\hat{P}(a)}{P_0(a)}$
- **Strong MAGIC:** $G \approx \chi^2(r-1)$ regardless of P_0!
- What is $\chi^2(d)$?

$$\chi^2(d) \sim \sum_{i=1}^{d} Z_i^2 \qquad Z_i \overset{iid}{\sim} \mathcal{N}(0,1)$$

$$\chi^2(d) = \texttt{scipy.stats.chi2.pdf}(\cdot, \texttt{df} = d)$$

Hacks: $\chi^2(d) \approx \mathcal{N}$ for $d \geq 500$ and $\sqrt{\chi^2} \approx \mathcal{N}$ for $d \geq 50$

- So the final test is: REJECT if $G > x_\alpha(r-1)$

Testing for discrete distribution (goodness-of-fit)

HT problem

- X be r-valued: $[r] = \{1, \ldots, r\}$
- P_0 – a pmf on $[r]$, i.e. $P_0(1) + \cdots + P_0(r) = 1$
- Null $H : X \overset{iid}{\sim} P_0$
- Alt $K : X \overset{iid}{\sim} P$ with $P \neq P_0$

G-test

- $\hat{P}(\cdot) = \frac{1}{n}\#\{j : x_j = \cdot\}$ empirical dist
- $g_{obs} = 2n \sum_{a=1}^{r} \hat{P}(a) \log \frac{\hat{P}(a)}{P_0(a)} = 2nD(\hat{P} \| P_0)$
- p-value= $\mathbb{P}[\chi^2(r-1) > g_{obs}]$

Remarks: Could use any other "distance" $d(\hat{P}, P_0)$ and simulate.

Social experiment

Rules of the game

- **Everyone** please think of **two** random bits
- Write them down!
- Now let me collect the results

Test 1: Generated bits are uniform coin flips?

- $n_0 = \#$ of 0 bits, $n_1 = \#$ of 1 bits.
- Calculate $G = 2n_0 \log \frac{2n_0}{n} + 2n_1 \log \frac{2n_1}{n}$
- Compare to quantiles of $\chi^2(1)$:

$P[\chi^2(1) > g]$	g
0.05	3.8
0.1	2.7
0.2	1.6
0.3	1.1

Test 2: Generated pairs of bits are (1/4, 1/2, 1/4)?

Testing for continuous distribution (goodness-of-fit)

- What if now null $H : X \overset{iid}{\sim} P_0$ with P_0 – continuous dist. on $\mathbb{R}$?
- For example: $P_0 = \mathcal{N}(0, 1)$?

Kolmogorov-Smirnov test

$$KS_n = \max_{-\infty < x < \infty} \sqrt{n} |\hat{F}_X(x) - F_0(x)|$$

- **MAGIC:** Distribution of KS_n is independent of P_0 (!!)

$KS_n >$ `scipy.stats.ksone.ppf`$(1 - \alpha,$ `n`$)$ then REJECT

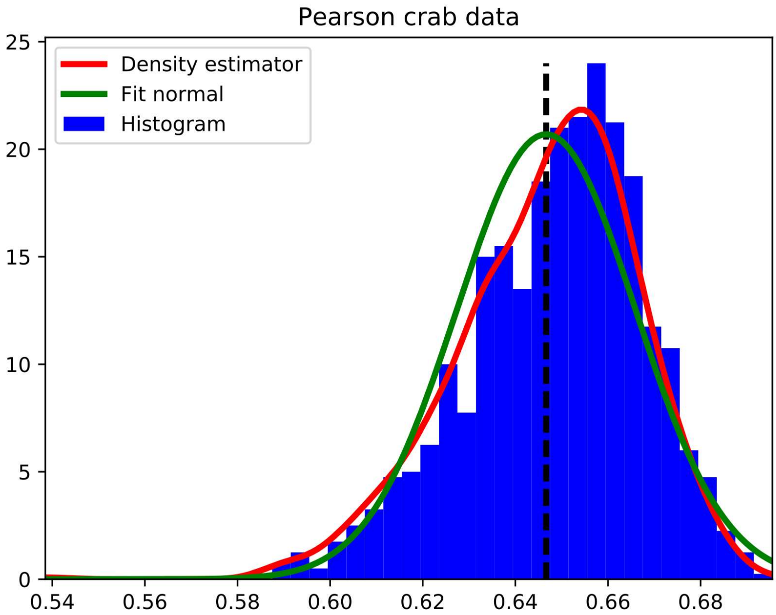
- Non-parametric stats.: dist. of KS_n is same for all P_0 in a huge class
- Don't trust `scipy`? Can do Monte Carlo with $P_0 = \mathrm{Uniform}[0, 1]$.
- For large n converges to explicit Kolmogorov distribution:

$$\mathbb{P}[KS_n \le x] \to \frac{\sqrt{2\pi}}{x} \sum_{k=1}^{\infty} e^{-(2k-1)^2 \pi^2 / (8x^2)}$$

- Example: Check if Pearson's crab data is normal.

Pearson crab data

BIN
0.5385
0.5875
0.5915
0.5955
0.5995
0.6035
0.6075
0.6115
0.6155
0.6195
0.6235
0.6275
0.6315
0.6355
0.6395



- Is
- N
- U
- G

Quantile-quantile plot

qq