Hypothesis testing 4

Y. Polyanskiy, D. Shah, J. Tsitsiklis

6.S077

2018

Outline:

- Recap (p-value, tests we learned)
- Type-1 and type-2 errors.
- Simple vs simple HT. Likelihood-ratio test. **Neyman-Pearson lemma**
- **Concept of power.**
- Most powerful tests. Unbiasedness.

Review

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:

- Find statistic $T(X_1, \ldots, X_n)$ with known dist. under H
- Test: $T > t_\alpha$ REJECT
- t_α is chosen depending on required size:

$$\max_{P \in \mathcal{C}_0} P[T > t_\alpha] \leq \alpha \,.$$

- Alternatively, report p-value: If $T(x_1, \ldots, x_n) = t_{obs}$

$$p = \max_{P \in \mathcal{C}_0} P[T > t_{obs}]$$

(aka "probability of same or more extreme data under null")

List of tests

We learned:

- 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 $X \not\sim \mathcal{N}(0,1)$
- 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 $X \not\perp\!\!\!\perp Y$

List of tests

We learned:

- 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 $X \not\perp\!\!\!\perp Y$
- TODAY: Only one more test (LRT).
  Of interest to: Bayesianists, theorists.

Curious asymmetry

- For each test we have only worried about

$$P[\text{REJECT}] \leq \alpha \qquad \forall P \in \mathcal{C}_0$$

Curious asymmetry

- For each test we have only worried about

$$P[\text{REJECT}] \leq \alpha \qquad \forall P \in \mathcal{C}_0$$

- In other words, we cared about type-1 error
- Definition: type-1 error happens when test rejects null, but null is true

Curious asymmetry

- For each test we have only worried about

$$P[\text{REJECT}] \leq \alpha \qquad \forall P \in \mathcal{C}_0$$

- In other words, we cared about type-1 error
- Definition: type-1 error happens when test rejects null, but null is true
- But if we are so worried about type-1 error only, why reject at all?
- Best test: whatever the data, always ACCEPT null
- What is bad about this test?

Curious asymmetry

- For each test we have only worried about

$$P[\text{REJECT}] \leq \alpha \qquad \forall P \in \mathcal{C}_0$$

- In other words, we cared about type-1 error
- Definition: type-1 error happens when test rejects null, but null is true
- But if we are so worried about type-1 error only, why reject at all?
- Best test: whatever the data, always ACCEPT null
- What is bad about this test? It has no power
- ... i.e. cannot detect alternative even if billions of samples are given.

Another mystery: Why we had K?

Definition

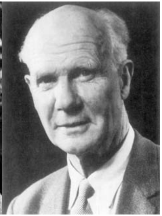
Statistical hypotheses:

- 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.

- We only talked about behavior under H. Why do we even have K?
- Fisher: We don't need to.

Another mystery: Why we had K?



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.

- We only talked about behavior under H. Why do we even have K?
- Fisher: We don't need to.
- Neyman-Pearson: need to control **power of test** to reject K
  (i.e. need to know what aspects of H are tested)
- Concept of **power** complicated in general

Easy (but important) special case

Simple vs simple HT

- null $H : X_i \overset{iid}{\sim} P_X$
- alt. $K : X_i \overset{iid}{\sim} Q_X$
- ... P_X and Q_X are known fixed distributions over some $\mathcal{X}$

Easy (but important) special case

Simple vs simple HT

- null $H : X_i \overset{iid}{\sim} P_X$
- alt. $K : X_i \overset{iid}{\sim} Q_X$
- ... P_X and Q_X are known fixed distributions over some $\mathcal{X}$

- When $|\mathcal{C}_0| = 1$ we say **null hypothesis is simple**
- When $|\mathcal{C}_1| = 1$ we say **alt. hypothesis is simple**
- ... so here we have simple vs. simple HT

Easy (but important) special case

Simple vs simple HT

- null $H : X_i \stackrel{iid}{\sim} P_X$
- alt. $K : X_i \stackrel{iid}{\sim} Q_X$
- ... P_X and Q_X are known fixed distributions over some $\mathcal{X}$

Easy (but important) special case

Simple vs simple HT

- null $H : X_i \stackrel{iid}{\sim} P_X$
- alt. $K : X_i \stackrel{iid}{\sim} Q_X$
- ... P_X and Q_X are known fixed distributions over some $\mathcal{X}$

- Any test is an algorithm

  if $(X_1, \ldots, X_n) \in E$ *then* **REJECT** *null.*

  So each set $E \subset \mathcal{X}^n$ determines a different test

Easy (but important) special case

Simple vs simple HT

- null $H : X_i \overset{iid}{\sim} P_X$
- alt. $K : X_i \overset{iid}{\sim} Q_X$
- ... P_X and Q_X are known fixed distributions over some $\mathcal{X}$

- Any test is an algorithm

  if $(X_1, \ldots, X_n) \in E$ *then* *REJECT* *null.*

  So each set $E \subset \mathcal{X}^n$ determines a different test
- How good is a given E? Need to compute 2x2 table:

	H true	K true
Test rejects	α	β
Test accepts	$1-\alpha$	$1-\beta$

- Type-1 error $= \alpha$. Type-2 error $= 1 - \beta$.

Example

- Suppose $\mathcal{X} = \{1, ..., 6\}$.
- Under null we have a fair die: $X \overset{iid}{\sim} P_X = [\frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}]$
- Under alt. we have a fair odd die: $X \overset{iid}{\sim} Q_X = [\frac{1}{3}, 0, \frac{1}{3}, 0, \frac{1}{3}, 0]$

Example

- Suppose $\mathcal{X} = \{1, ..., 6\}$.
- Under null we have a fair die: $X \overset{iid}{\sim} P_X = [\frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}]$
- Under alt. we have a fair odd die: $X \overset{iid}{\sim} Q_X = [\frac{1}{3}, 0, \frac{1}{3}, 0, \frac{1}{3}, 0]$
- Consider test: If $T = \frac{1}{n} \sum_{i=1}^{n} X_i \leq 3.25$ then REJECT
- For $n = 1$ we have:

$$P_X[T \leq 3.25] = \frac{1}{2}, \qquad Q_X[T \leq 3.25] = \frac{2}{3}$$

Example

- Suppose $\mathcal{X} = \{1, ..., 6\}$.
- Under null we have a fair die: $X \overset{iid}{\sim} P_X = [\frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}]$
- Under alt. we have a fair odd die: $X \overset{iid}{\sim} Q_X = [\frac{1}{3}, 0, \frac{1}{3}, 0, \frac{1}{3}, 0]$
- Consider test: If $T = \frac{1}{n}\sum_{i=1}^{n} X_i \leq 3.25$ then REJECT
- For $n = 1$ we have:

$$P_X[T \leq 3.25] = \frac{1}{2}, \qquad Q_X[T \leq 3.25] = \frac{2}{3}$$

$n = 1$

	H true	K true
Test rejects	0.5	0.67
Test accepts	0.5	0.33

Example

- Suppose $\mathcal{X} = \{1, ..., 6\}$.
- Under null we have a fair die: $X \overset{iid}{\sim} P_X = [\frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}]$
- Under alt. we have a fair odd die: $X \overset{iid}{\sim} Q_X = [\frac{1}{3}, 0, \frac{1}{3}, 0, \frac{1}{3}, 0]$
- Consider test: If $T = \frac{1}{n}\sum_{i=1}^{n} X_i \le 3.25$ then REJECT
- For $n = 1$ we have:

$$P_X[T \le 3.25] = \frac{1}{2}, \qquad Q_X[T \le 3.25] = \frac{2}{3}$$

$n = 2$

	H true	K true
Test rejects	0.42	0.67
Test accepts	0.58	0.33

Example

- Suppose $\mathcal{X} = \{1, ..., 6\}$.
- Under null we have a fair die: $X \overset{iid}{\sim} P_X = [\frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}]$
- Under alt. we have a fair odd die: $X \overset{iid}{\sim} Q_X = [\frac{1}{3}, 0, \frac{1}{3}, 0, \frac{1}{3}, 0]$
- Consider test: If $T = \frac{1}{n}\sum_{i=1}^{n} X_i \leq 3.25$ then REJECT
- For $n = 1$ we have:

$$P_X[T \leq 3.25] = \frac{1}{2}, \qquad Q_X[T \leq 3.25] = \frac{2}{3}$$

$n = 10$

	H true	K true
Test rejects	0.32	0.71
Test accepts	0.68	0.29

Example

- Suppose $\mathcal{X} = \{1, ..., 6\}$.
- Under null we have a fair die: $X \overset{iid}{\sim} P_X = [\frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}]$
- Under alt. we have a fair odd die: $X \overset{iid}{\sim} Q_X = [\frac{1}{3}, 0, \frac{1}{3}, 0, \frac{1}{3}, 0]$
- Consider test: If $T = \frac{1}{n}\sum_{i=1}^{n} X_i \leq 3.25$ then REJECT
- For $n = 1$ we have:

$$P_X[T \leq 3.25] = \frac{1}{2}, \qquad Q_X[T \leq 3.25] = \frac{2}{3}$$

$$n = 50$$

	H true	K true
Test rejects	0.15	0.87
Test accepts	0.85	0.13

Example

- Suppose $\mathcal{X} = \{1, ..., 6\}$.
- Under null we have a fair die: $X \overset{iid}{\sim} P_X = [\frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}]$
- Under alt. we have a fair odd die: $X \overset{iid}{\sim} Q_X = [\frac{1}{3}, 0, \frac{1}{3}, 0, \frac{1}{3}, 0]$
- Consider test: If $T = \frac{1}{n}\sum_{i=1}^{n} X_i \leq 3.25$ then REJECT
- For $n = 1$ we have:

$$P_X[T \leq 3.25] = \frac{1}{2}, \qquad Q_X[T \leq 3.25] = \frac{2}{3}$$

$n = 150$

	H true	K true
Test rejects	0.04	0.97
Test accepts	0.96	0.03

Example

- Suppose $\mathcal{X} = \{1, ..., 6\}$.
- Under null we have a fair die: $X \overset{iid}{\sim} P_X = [\frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}]$
- Under alt. we have a fair odd die: $X \overset{iid}{\sim} Q_X = [\frac{1}{3}, 0, \frac{1}{3}, 0, \frac{1}{3}, 0]$
- Consider test: If $T = \frac{1}{n}\sum_{i=1}^{n} X_i \leq 3.25$ then REJECT
- For $n = 1$ we have:

$$P_X[T \leq 3.25] = \frac{1}{2}, \qquad Q_X[T \leq 3.25] = \frac{2}{3}$$

$$n = 150$$

	H true	K true
Test rejects	0.04	0.97
Test accepts	0.96	0.03

- So with n we get $\alpha \to 0$ and $\beta \to 1$ (this is typical)

Example

- Suppose $\mathcal{X} = \{1, ..., 6\}$.
- Under null we have a fair die: $X \overset{iid}{\sim} P_X = [\frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}]$
- Under alt. we have a fair odd die: $X \overset{iid}{\sim} Q_X = [\frac{1}{3}, 0, \frac{1}{3}, 0, \frac{1}{3}, 0]$
- Consider test: If $T = \frac{1}{n}\sum_{i=1}^{n} X_i \leq 3.25$ then REJECT
- For $n = 1$ we have:

$$P_X[T \leq 3.25] = \frac{1}{2}, \qquad Q_X[T \leq 3.25] = \frac{2}{3}$$

$n = 150$

	H true	K true
Test rejects	0.04	0.97
Test accepts	0.96	0.03

- So with n we get $\alpha \to 0$ and $\beta \to 1$ (this is typical)
- Usually, $\beta = 1$ is not possible (requires $\alpha = 1$).
  But in this case, you **can** get $\beta = 1$ and $\alpha \to 0$. How?

Example

- Suppose $\mathcal{X} = \{1, ..., 6\}$.
- Under null we have a fair die: $X \overset{iid}{\sim} P_X = [\frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}]$
- Under alt. we have a fair odd die: $X \overset{iid}{\sim} Q_X = [\frac{1}{3}, 0, \frac{1}{3}, 0, \frac{1}{3}, 0]$
- Consider test: If $T = \frac{1}{n}\sum_{i=1}^{n} X_i \leq 3.25$ then REJECT
- For $n = 1$ we have:

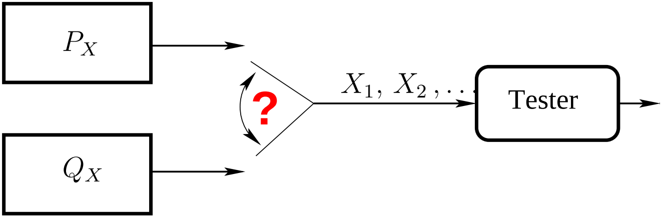
$$P_X[T \leq 3.25] = \frac{1}{2}, \qquad Q_X[T \leq 3.25] = \frac{2}{3}$$

$n = 150$

	H true	K true
Test rejects	0.04	0.97
Test accepts	0.96	0.03

- So with n we get $\alpha \to 0$ and $\beta \to 1$ (this is typical)
- Usually, $\beta = 1$ is not possible (requires $\alpha = 1$).
  But in this case, you **can** get $\beta = 1$ and $\alpha \to 0$. How?
- Next: **How do we find better (and less ad hoc) tests?**

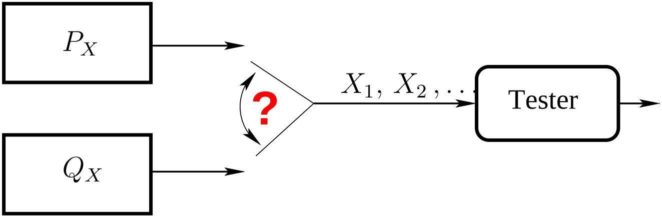
Easy case: Two simple hypotheses



- Return to general case. Test = rejection set $E \subset \mathcal{X}^n$
- For each test:

$$\begin{aligned} \text{size} &= \alpha \triangleq \mathbb{P}[\text{REJECT}] \\ \text{power} &= \beta \triangleq \mathbb{Q}[\text{REJECT}] \end{aligned}$$

Easy case: Two simple hypotheses



- Return to general case. Test = rejection set $E \subset \mathcal{X}^n$
- For each test:

$$\begin{aligned} \text{size} &= \alpha \triangleq \mathbb{P}[\text{REJECT}] \\ \text{power} &= \beta \triangleq \mathbb{Q}[\text{REJECT}] \end{aligned}$$

- if tests never rejects: $\alpha = 0$, $\beta = 0$ (i.e. $E = \emptyset$)
- randomized test: reject w.p. α w/o looking at the data: $\beta = \alpha$
- Best tradeoff?

Solve binary HT $\iff$ maximize β subject to α

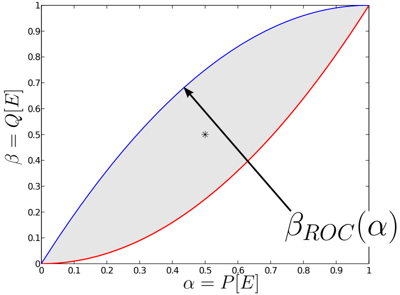
Neyman-Pearson's ROC curve

Definition

Let P and Q be distributions of $(X_1, \ldots, X_n)$ under null and alt. Then ROC-curve is

$$\beta_{ROC}(\alpha) \triangleq \max_{E: P[E] \le \alpha} Q[E] \,.$$

iterate over
all "sets" E $\Rightarrow$
plot pairs $(P[E], Q[E])$



- **Meaning:** $\beta_{ROC}(\alpha) =$ power of best size-α test.
- $\alpha =$false-positive rate, $\beta =$ true-positive rate
- Later: in logistic regression and binary classification.

How to compute $\beta(\alpha)$?

Theorem (Neyman-Pearson)

Fix any $0 \le \gamma \le +\infty$*. Consider the test*

$$E = \left\{ X : \frac{P(X)}{Q(X)} \le \gamma \right\}$$

It achieves $\alpha = P[E]$ *and* $\beta = Q[E]$ *that belong to the boundary curve* $\beta_{ROC}(\alpha)$*. As* γ *varies these tests trace all of* $\beta_{ROC}(\alpha)$*.*

- Thus, likelihood-ratio tests are OPTIMAL. For iid data:

How to compute $\beta(\alpha)$?

Theorem (Neyman-Pearson)

Fix any $0 \le \gamma \le +\infty$. Consider the test

$$E = \left\{ X : \frac{P(X)}{Q(X)} \le \gamma \right\}$$

It achieves $\alpha = P[E]$ and $\beta = Q[E]$ that belong to the boundary curve $\beta_{ROC}(\alpha)$. As γ varies these tests trace all of $\beta_{ROC}(\alpha)$.

- Thus, likelihood-ratio tests are OPTIMAL. For iid data:

Optimal test for simple HT

$$T = -\sum_{i=1}^{n} \log \frac{P(X_i)}{Q(X_i)}$$

If $T \ge t$ REJECT

- Selecting threshold: use simulations or normal approximation ($\mathbb{E}[T] = nD(P\|Q)$, $\mathrm{Var}[T] = \cdots$)

How to compute $\beta(\alpha)$?

Theorem (Neyman-Pearson)

Fix any $0 \le \gamma \le +\infty$. *Consider the test*

$$E = \left\{ X : \frac{P(X)}{Q(X)} \le \gamma \right\}$$

It achieves $\alpha = P[E]$ *and* $\beta = Q[E]$ *that belong to the boundary curve* $\beta_{ROC}(\alpha)$. *As* γ *varies these tests trace all of* $\beta_{ROC}(\alpha)$.

- Thus, likelihood-ratio tests are OPTIMAL. For iid data:

Optimal test for simple HT

$$T = -\sum_{i=1}^{n} \log \frac{P(X_i)}{Q(X_i)}$$

If $T \ge t$ REJECT

- Selecting threshold: use simulations or normal approximation ($\mathbb{E}[T] = nD(P\|Q)$, $\mathrm{Var}[T] = \cdots$)
- Stein's lemma: $\beta(\alpha) = e^{-nD(P\|Q)+o(n)}$

Example: Detection in communication systems

Binary detection in Gaussian noise

- Null H: $X_i \sim \mathcal{N}(0,1)$ (white noise)
- Alt K: $X_i \sim \mathcal{N}(a_i,1)$ (signal + noise)
- Goal: decide between the two

- Write-down LLR:

$$T = -\sum_i \log \frac{P(X_i)}{Q(X_i)} = C + \sum_{i=1}^n a_i x_i = C + \boldsymbol{a} \cdot \boldsymbol{x}$$

where C is some constant, indep. of $\boldsymbol{x}$. So take $C = 0$.

Example: Detection in communication systems

Binary detection in Gaussian noise

- Null H: $X_i \sim \mathcal{N}(0,1)$ (white noise)
- Alt K: $X_i \sim \mathcal{N}(a_i,1)$ (signal + noise)
- Goal: decide between the two

- Write-down LLR:

$$T = -\sum_i \log \frac{P(X_i)}{Q(X_i)} = C + \sum_{i=1}^n a_i x_i = C + \boldsymbol{a} \cdot \boldsymbol{x}$$

  where C is some constant, indep. of $\boldsymbol{x}$. So take $C = 0$.
- Under null: $T \sim \mathcal{N}(0, \|\boldsymbol{a}\|^2)$.

Example: Detection in communication systems

Binary detection in Gaussian noise

- Null H: $X_i \sim \mathcal{N}(0,1)$ (white noise)
- Alt K: $X_i \sim \mathcal{N}(a_i,1)$ (signal + noise)
- Goal: decide between the two

- Write-down LLR:

$$T = -\sum_i \log \frac{P(X_i)}{Q(X_i)} = C + \sum_{i=1}^n a_i x_i = C + \boldsymbol{a} \cdot \boldsymbol{x}$$

  where C is some constant, indep. of $\boldsymbol{x}$. So take $C = 0$.
- Under null: $T \sim \mathcal{N}(0, \|\boldsymbol{a}\|^2)$. Under alt.: $T \sim \mathcal{N}(\|\boldsymbol{a}\|^2, \|\boldsymbol{a}\|^2)$

Example: Detection in communication systems

Binary detection in Gaussian noise

- Null H: $X_i \sim \mathcal{N}(0,1)$ (white noise)
- Alt K: $X_i \sim \mathcal{N}(a_i,1)$ (signal + noise)
- Goal: decide between the two

- Write-down LLR:

$$T = -\sum_i \log \frac{P(X_i)}{Q(X_i)} = C + \sum_{i=1}^n a_i x_i = C + \boldsymbol{a} \cdot \boldsymbol{x}$$

  where C is some constant, indep. of $\boldsymbol{x}$. So take $C = 0$.
- Under null: $T \sim \mathcal{N}(0, \|\boldsymbol{a}\|^2)$. Under alt.: $T \sim \mathcal{N}(\|\boldsymbol{a}\|^2, \|\boldsymbol{a}\|^2)$
- Thus, optimal decoder is:

$$\frac{\mathbf{a} \cdot \mathbf{x}}{\|\mathbf{a}\|} > z_\alpha \quad \Rightarrow \quad \boxed{\text{DETECT ALARM}}$$

- Its ROC curve – see PSET.

Example: Detection in communication systems

Binary detection in Gaussian noise

- Null H: $X_i \sim \mathcal{N}(0,1)$ (white noise)
- Alt K: $X_i \sim \mathcal{N}(a_i,1)$ (signal + noise)
- Goal: decide between the two

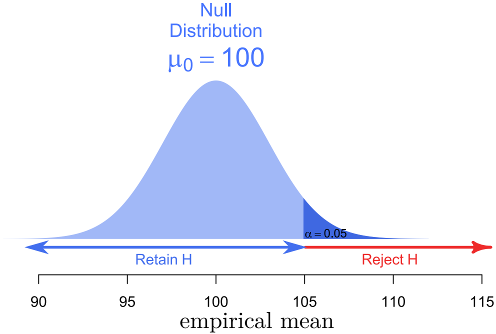
- Write-down LLR:

$$T = -\sum_i \log \frac{P(X_i)}{Q(X_i)} = C + \sum_{i=1}^n a_i x_i = C + \boldsymbol{a}\cdot\boldsymbol{x}$$

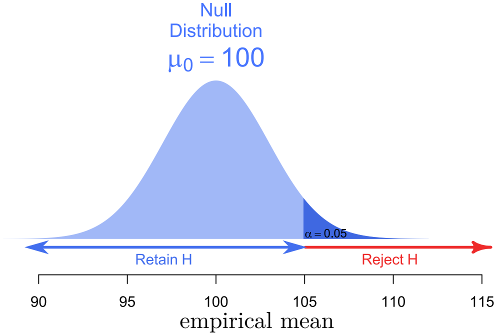
  where C is some constant, indep. of $\boldsymbol{x}$. So take $C=0$.
- Under null: $T \sim \mathcal{N}(0, \|\boldsymbol{a}\|^2)$. Under alt.: $T \sim \mathcal{N}(\|\boldsymbol{a}\|^2, \|\boldsymbol{a}\|^2)$
- Thus, optimal decoder is:

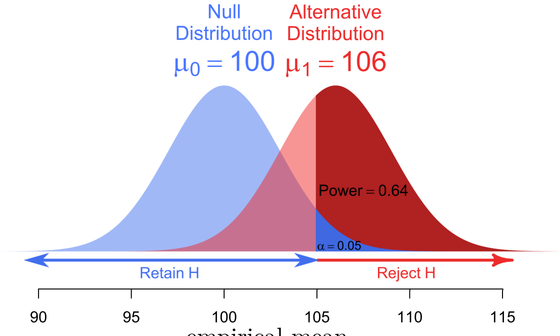
$$\frac{\mathbf{a}\cdot\mathbf{x}}{\|\mathbf{a}\|} > z_\alpha \quad \Rightarrow \quad \boxed{\text{DETECT ALARM}}$$

- Its ROC curve – see PSET.
- Pictorially, distribution of T is...



Null
Distribution
$\mu_0 = 100$
$\alpha = 0.05$
Retain H
Reject H
90
95
100
105
110
115
empirical mean


Null
Distribution
$\mu_0 = 100$
$\alpha = 0.05$
Retain H
Reject H
90
95
100
105
110
115
empirical mean


Null
Distribution
$\mu_0 = 100$
Alternative
Distribution
$\mu_1 = 106$
Power = 0.64
$\alpha = 0.05$
Retain H
Reject H
90
95
100
105
110
115

General definition of power

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 $P \in \mathcal{C}_1$

- Fix some test (i.e. a rejection degion), e.g. $\{T(X) > t\}$

General definition of power

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 $P \in \mathcal{C}_1$

- Fix some test (i.e. a rejection degion), e.g. $\{T(X) > t\}$

Definition (Power of a test)

Power is a function of distribution P

$$\beta(P) = P[\text{test rejects null}]$$

- As $P \in \mathcal{C}_1$ varies, so does the power of test to correctly reject null.
- Recall that test is of size-α

$$\beta(P) \leq \alpha \qquad \forall P \in \mathcal{C}_0$$

Sometimes, only have $\leq \alpha + o(1)$ as $n \to \infty$

General definition of power

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 $P \in \mathcal{C}_1$

- Fix some test (i.e. a rejection degion), e.g. $\{T(X) > t\}$

Definition (Power of a test)

Power is a function of distribution P

$$\beta(P) = P[\text{test rejects null}]$$

- As $P \in \mathcal{C}_1$ varies, so does the power of test to correctly reject null.
- Recall that test is of size-α

$$\beta(P) \leq \alpha \qquad \forall P \in \mathcal{C}_0$$

Sometimes, only have $\leq \alpha + o(1)$ as $n \to \infty$

- Let's illustrate on example of parametric tests

Reminder: Wald test for parametric models

- Have a good parametric model $X \overset{iid}{\sim} P_\theta$ (e.g. $X \sim \mathrm{Ber}(\theta)$)

Reminder: Wald test for parametric models

- Have a good parametric model $X \overset{iid}{\sim} P_\theta$ (e.g. $X \sim \mathrm{Ber}(\theta)$)
- **Special case:** θ – scalar, $\Theta_0 = \{\theta_0\}$:

$$H_0 : \theta = \theta_0 \qquad \text{vs.} \qquad H_1 : \theta \neq \theta_0 \,.$$

The Wald test-statistic

$$W = \frac{\hat{\theta} - \theta_0}{\widehat{se}}$$

$\hat{\theta}$ = parameter estimator (eg. MLE),
$\widehat{se}$ = estimate of std.err. (e.g. $\widehat{se}^2 = \frac{1}{n-1}\hat{\theta}(1-\hat{\theta})$)

- Assuming a) asymptotic normality of $\hat{\theta}$, b) consistency of $\widehat{se}^2$:

$$W \approx \mathcal{N}(0,1) \qquad \text{for large } n$$

- So the Wald test (two-sided): If $|W| > z_{\frac{\alpha}{2}}$ then REJECT.

Reminder: Wald test for parametric models

- Have a good parametric model $X \overset{iid}{\sim} P_\theta$ (e.g. $X \sim \mathrm{Ber}(\theta)$)
- **Special case:** θ – scalar, $\Theta_0 = \{\theta_0\}$:

$$H_0 : \theta = \theta_0 \qquad \text{vs.} \qquad H_1 : \theta \neq \theta_0 \,.$$

The Wald test-statistic

$$W = \frac{\hat{\theta} - \theta_0}{\widehat{se}}$$

$\hat{\theta}$ = parameter estimator (eg. MLE),
$\widehat{se}$ = estimate of std.err. (e.g. $\widehat{se}^2 = \frac{1}{n-1}\hat{\theta}(1-\hat{\theta})$)

- Assuming a) asymptotic normality of $\hat{\theta}$, b) consistency of $\widehat{se}^2$:

$$W \approx \mathcal{N}(0,1) \qquad \text{for large } n$$

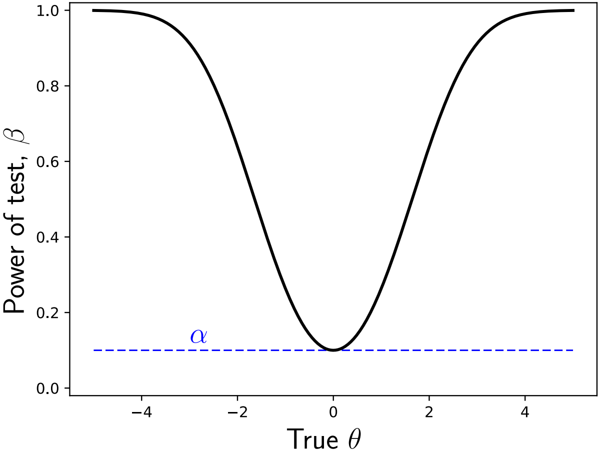
- So the Wald test (two-sided): If $|W| > z_{\frac{\alpha}{2}}$ then REJECT.
- Once threshold is fixed can compute power $\beta(\theta)$ as a function of θ

Illustration of power behavior

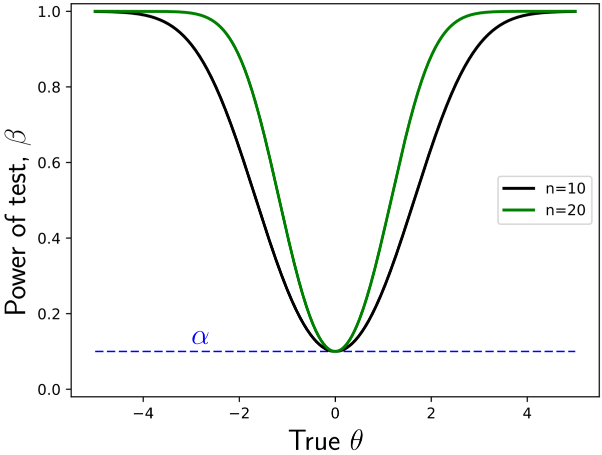
Open: `figs/zEffectSize.gif` in browser

Orig. URL:
`http://my.ilstu.edu/~wjschne/138/Psychology138Lab14.html`

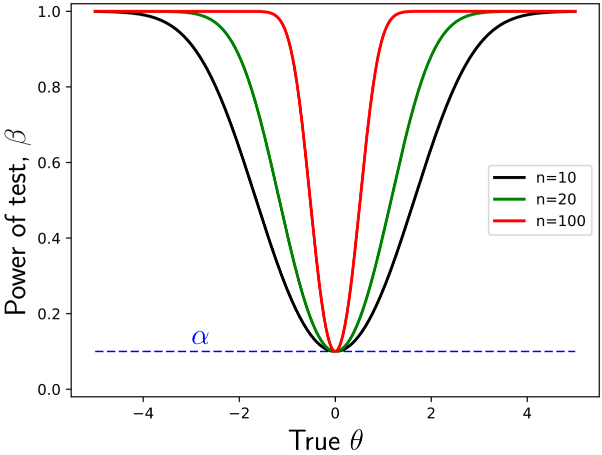
Power as a function of θ ... and number of samples



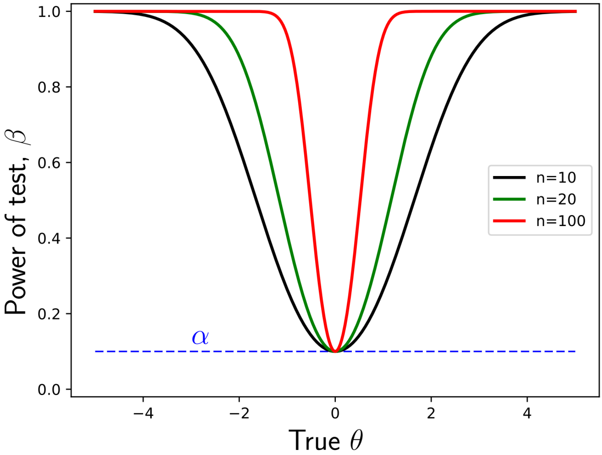
Power as a function of θ ... and number of samples



Power as a function of θ ... and number of samples



Power as a function of θ ... and number of samples



- Tests need to be properly powered
- expert (doctor/engineer/etc) needs to specify resolution of θ
- statistician determines $\#$ of samples required to resolve 0 from $\neq 0$.

Example of under-powered test

$H:$ *Rain in Seattle is iid w.p. $1/2$ everyday*

- $\mu_0 = 1/2$, $\sigma_0^2 = 1/4$. Test of size $\alpha = 0.05$ will be:

REJECT null if $|Z| > 1.96$

Example of under-powered test

$H:$ *Rain in Seattle is iid w.p.* $1/2$ *everyday*

- $\mu_0 = 1/2$, $\sigma_0^2 = 1/4$. Test of size $\alpha = 0.05$ will be:

REJECT null if $|Z| > 1.96$

- Take random n days from Seattle weather data

DATE	RAIN
1950-05-25	False
1957-07-14	True
1960-05-18	False
1982-02-10	False
1994-08-11	False
1994-09-24	False
1996-07-29	False
1997-06-01	True
1997-12-22	True
2001-02-14	False

$Z = -1.26$

Example of under-powered test

H: Rain in Seattle is iid w.p. $1/2$ *everyday*

- $\mu_0 = 1/2$, $\sigma_0^2 = 1/4$. Test of size $\alpha = 0.05$ will be:

REJECT null if $|Z| > 1.96$

- Take random n days from Seattle weather data
- $n = 10$ samples, got Z-scores:

$$Z = 0.0, -1.89, -0.63, 0.63, -0.63$$

Test: accept null

DATE	RAIN
1950-05-25	False
1957-07-14	True
1960-05-18	False
1982-02-10	False
1994-08-11	False
1994-09-24	False
1996-07-29	False
1997-06-01	True
1997-12-22	True
2001-02-14	False

$Z = -1.26$

Example of under-powered test

H: Rain in Seattle is iid w.p. $1/2$ everyday

- $\mu_0 = 1/2$, $\sigma_0^2 = 1/4$. Test of size $\alpha = 0.05$ will be:

REJECT null if $|Z| > 1.96$

- Take random n days from Seattle weather data
- $n = 10$ samples, got Z-scores:

$$Z = 0.0, -1.89, -0.63, 0.63, -0.63$$

Test: accept null

- $n = 100$, got Z-scores:

$$Z = -3.2, -2.0, -1.4, -3.6, -1.2$$

Test: sometimes reject sometimes accept

DATE	RAIN
1950-05-25	False
1957-07-14	True
1960-05-18	False
1982-02-10	False
1994-08-11	False
1994-09-24	False
1996-07-29	False
1997-06-01	True
1997-12-22	True
2001-02-14	False

$Z = -1.26$

Example of under-powered test

$H:$ *Rain in Seattle is iid w.p.* $1/2$ *everyday*

- $\mu_0 = 1/2$, $\sigma_0^2 = 1/4$. Test of size $\alpha = 0.05$ will be:

REJECT null if $|Z| > 1.96$

- Take random n days from Seattle weather data
- $n = 10$ samples, got Z-scores:

$$Z = 0.0, -1.89, -0.63, 0.63, -0.63$$

Test: accept null

- $n = 100$, got Z-scores:

$$Z = -3.2, -2.0, -1.4, -3.6, -1.2$$

Test: sometimes reject sometimes accept

- $n = 1000$, got Z-scores:

$$Z = -5.76, -2.72, -3.29, -3.92, -4.42$$

Test: reject null (always)

(FYI: over 1948-2017 $\hat{P}[\text{rain}] = 42.6\%$ $\leftarrow$ TRUE θ)

DATE	RAIN
1950-05-25	False
1957-07-14	True
1960-05-18	False
1982-02-10	False
1994-08-11	False
1994-09-24	False
1996-07-29	False
1997-06-01	True
1997-12-22	True
2001-02-14	False

$Z = -1.26$

Comparing tests

- We now want to see how tests compare.
- The general idea:
  - Consider two tests (e.g. z-test vs G-test)
  - Choose thresholds in both so that size=α
  - Compute $\beta(\theta)$ for both tests
  - see if one is above the other.

Comparing tests

- We now want to see how tests compare.
- The general idea:
    - Consider two tests (e.g. z-test vs G-test)
    - Choose thresholds in both so that size=α
    - Compute $\beta(\theta)$ for both tests
    - see if one is above the other.
- Hard task in general. So let us consider very toy (parametric) models.

Gaussian location model

- We focus on one special parametric model
- $X \overset{iid}{\sim} \mathcal{N}(\theta, 1)$
- known-variance, Gaussian location model **(GLM)**

Gaussian location model

- We focus on one special parametric model
- $X \stackrel{iid}{\sim} \mathcal{N}(\theta, 1)$
- known-variance, Gaussian location model **(GLM)**
- Our null is always simple $\theta = 0$.

Gaussian location model

- We focus on one special parametric model
- $X \overset{iid}{\sim} \mathcal{N}(\theta, 1)$
- known-variance, Gaussian location model **(GLM)**
- Our null is always simple $\theta = 0$.
- Alternatives:

$$H : \theta = 0 \qquad \text{vs.} \qquad K : \theta > 0 \quad \text{(one-sided)}$$

$$H : \theta = 0 \qquad \text{vs.} \qquad K : \theta \neq 0 \quad \text{(two-sided)}$$

- So $\beta(\theta)$ is a function of a scalar θ

Gaussian location model

- We focus on one special parametric model
- $X \overset{iid}{\sim} \mathcal{N}(\theta, 1)$
- known-variance, Gaussian location model **(GLM)**
- Our null is always simple $\theta = 0$.
- Alternatives:

$$H : \theta = 0 \qquad \text{vs.} \qquad K : \theta > 0 \quad \text{(one-sided)}$$

$$H : \theta = 0 \qquad \text{vs.} \qquad K : \theta \neq 0 \quad \text{(two-sided)}$$

- So $\beta(\theta)$ is a function of a scalar θ
- Test is size-α iff $\beta(0) = \alpha$

One-sided test has more power

$$H : \theta = 0 \qquad \text{vs.} \qquad K : \theta > 0$$

In such cases we recommended one-sided tests:

$$Z > z_\alpha \quad \Rightarrow \quad \boxed{\text{REJECT}}$$

One-sided test has more power

$$H : \theta = 0 \qquad \text{vs.} \qquad K : \theta > 0$$

In such cases we recommended one-sided tests:

$$Z > z_\alpha \quad \Rightarrow \quad \text{REJECT}$$

Why not use two-sided test:

$$|Z| > z_{\frac{\alpha}{2}} \quad \Rightarrow \quad \text{REJECT}$$

One-sided test has more power

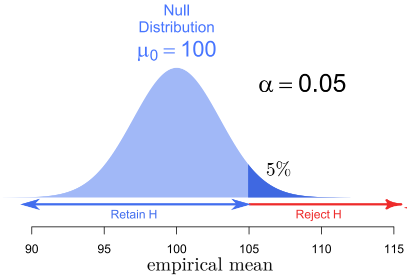
$$H : \theta = 0 \qquad \text{vs.} \qquad K : \theta > 0$$

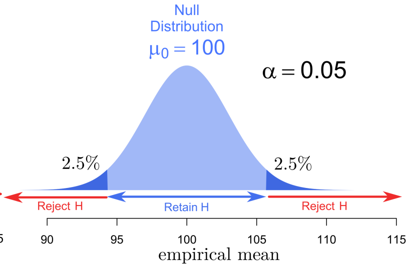
In such cases we recommended one-sided tests:

$$Z > z_\alpha \quad \Rightarrow \quad \boxed{\text{REJECT}}$$

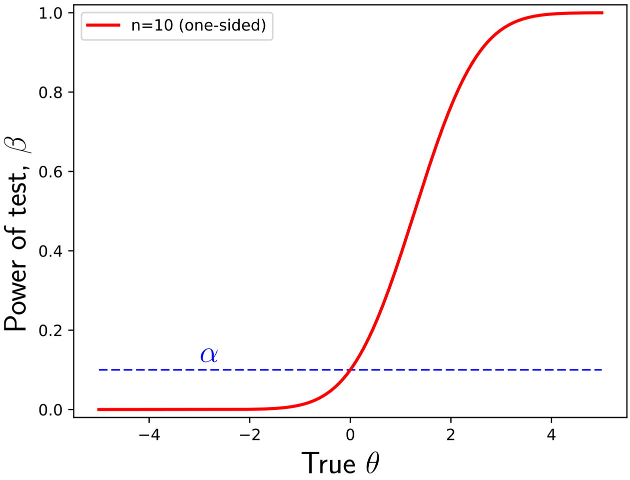
Why not use two-sided test:

$$|Z| > z_{\frac{\alpha}{2}} \quad \Rightarrow \quad \boxed{\text{REJECT}}$$

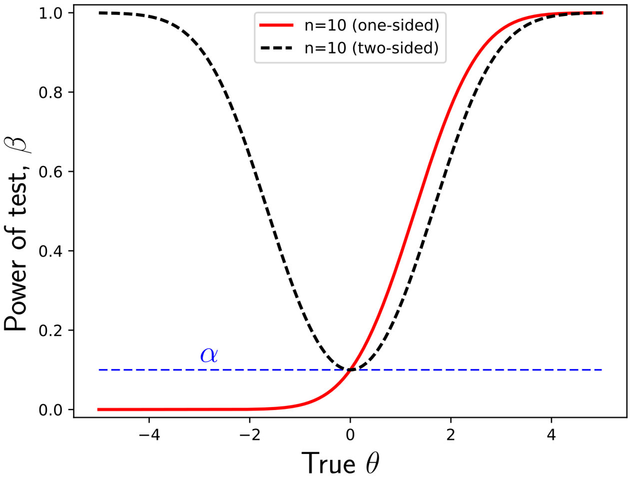




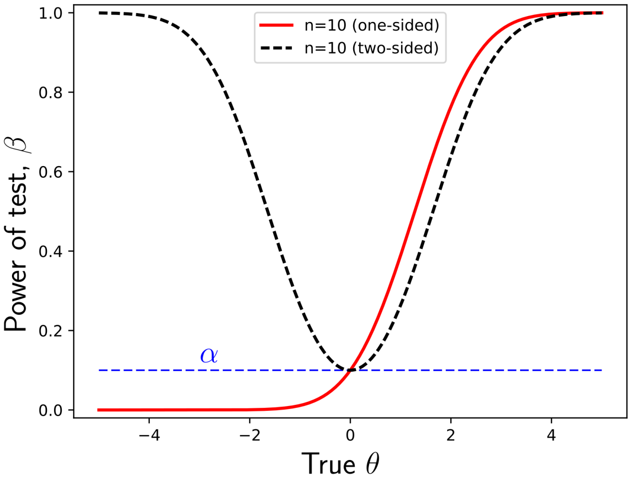
One-sided test has more power



One-sided test has more power



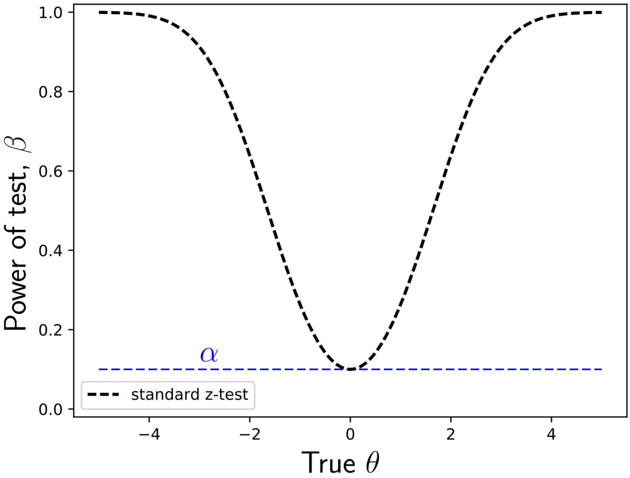
One-sided test has more power



- Punchline: Since $\theta \geq 0$ by assumption, one-sided test is strictly better.

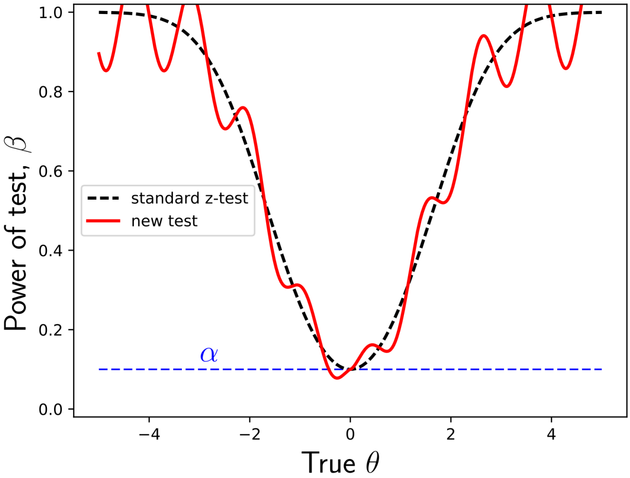
How to compare two tests in general?

How to compare two tests in general?



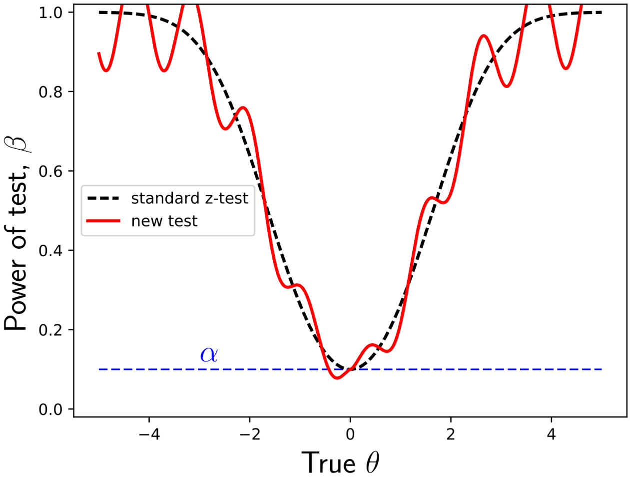
- MIT student proposes a new test. Is it better than standard?

How to compare two tests in general?



- MIT student proposes a new test. Is it better than standard?
- Is test size-α? Yes.

How to compare two tests in general?



- MIT student proposes a new test. Is it better than standard?
- Is test size-α? Yes. Is it better than the standard? Hmm...

The (doomed) quest for optimality

- Consider two size-α tests T and T' with powers $\beta_T(\theta)$ and $\beta_{T'}(\theta)$.
- We can definitely say that one is better than the other if

$$\beta_T(\theta) \geq \beta_{T'}(\theta) \qquad \forall \theta \in K$$

($\theta \in K$ means all θ in alternative hypotheses)

The (doomed) quest for optimality

- Consider two size-α tests T and T' with powers $\beta_T(\theta)$ and $\beta_{T'}(\theta)$.
- We can definitely say that one is better than the other if

$$\beta_T(\theta) \geq \beta_{T'}(\theta) \qquad \forall \theta \in K$$

  ($\theta \in K$ means all θ in alternative hypotheses)
- One would want a size-α test with the property:

$$\beta_T(\theta) \geq \beta_{T'}(\theta) \qquad \forall \text{size-}\alpha \text{ tests } T'$$

The (doomed) quest for optimality

- Consider two size-α tests T and T' with powers $\beta_T(\theta)$ and $\beta_{T'}(\theta)$.
- We can definitely say that one is better than the other if

$$\beta_T(\theta) \geq \beta_{T'}(\theta) \qquad \forall \theta \in K$$

($\theta \in K$ means all θ in alternative hypotheses)

- One would want a size-α test with the property:

$$\beta_T(\theta) \geq \beta_{T'}(\theta) \qquad \forall \text{size-}\alpha \text{ tests } T'$$

- If exists, such a test is called **most powerful**.
(if it is also unique, then we say T is **UMP**).

The (doomed) quest for optimality

- Consider two size-α tests T and T' with powers $\beta_T(\theta)$ and $\beta_{T'}(\theta)$.
- We can definitely say that one is better than the other if

$$\beta_T(\theta) \geq \beta_{T'}(\theta) \qquad \forall \theta \in K$$

  ($\theta \in K$ means all θ in alternative hypotheses)
- One would want a size-α test with the property:

$$\boxed{\beta_T(\theta) \geq \beta_{T'}(\theta) \qquad \forall \text{size-}\alpha \text{ tests } T'}$$

- If exists, such a test is called **most powerful**.
  (if it is also unique, then we say T is **UMP**).
- Sadly: Most of the time it does not exist.
- ... That is why HT is an art and many hundreds of tests exist.

The (doomed) quest for optimality

- Consider two size-α tests T and T' with powers $\beta_T(\theta)$ and $\beta_{T'}(\theta)$.
- We can definitely say that one is better than the other if

$$\beta_T(\theta) \geq \beta_{T'}(\theta) \qquad \forall \theta \in K$$

  ($\theta \in K$ means all θ in alternative hypotheses)
- One would want a size-α test with the property:

$$\beta_T(\theta) \geq \beta_{T'}(\theta) \qquad \forall \text{size-}\alpha \text{ tests } T'$$

- If exists, such a test is called **most powerful**.
  (if it is also unique, then we say T is **UMP**).
- Sadly: Most of the time it does not exist.
- ... That is why HT is an art and many hundreds of tests exist.
- However, in special cases they do exist:
  - One-sided GLM: UMP exists
  - Two-sided GLM: no UMP, but $\exists$ UMP-unbiased (**UMPU**)
  - A list of UMPU

Bayesian intermission

Bayesian resolution of "optimal test" problem

- Recall: Bayesianist has a prior for everything
- So he says: **Under null H, data is generated as:**
  1. Nature picks $\theta \in \mathcal{C}_0$ via $\theta \sim P^{(H)}$
     ($P_\theta^{(H)}$ is the secret sauce: the magic prior on θ's)

Bayesian intermission

Bayesian resolution of "optimal test" problem

- Recall: Bayesianist has a prior for everything
- So he says: **Under null H, data is generated as:**
  1. Nature picks $\theta \in \mathcal{C}_0$ via $\theta \sim P^{(H)}$
     ($P_\theta^{(H)}$ is the secret sauce: the magic prior on θ's)
  2. Generate $X_i \overset{iid}{\sim} P^\theta$

Bayesian intermission

Bayesian resolution of "optimal test" problem

- Recall: Bayesianist has a prior for everything
- So he says: **Under null H, data is generated as:**
  1. Nature picks $\theta \in \mathcal{C}_0$ via $\theta \sim P^{(H)}$
     ($P_\theta^{(H)}$ is the secret sauce: the magic prior on θ's)
  2. Generate $X_i \overset{iid}{\sim} P^\theta$
- (similar under alt. $K : \theta \sim P^{(K)}$ etc.

Bayesian intermission

Bayesian resolution of "optimal test" problem

- Recall: Bayesianist has a prior for everything
- So he says: **Under null H, data is generated as:**
  1. Nature picks $\theta \in \mathcal{C}_0$ via $\theta \sim P^{(H)}$
     ($P_\theta^{(H)}$ is the secret sauce: the magic prior on θ's)
  2. Generate $X_i \overset{iid}{\sim} P^\theta$
- (similar under alt. $K : \theta \sim P^{(K)}$ etc.
- This is awesome: now we have a simple vs simple HT:

$$H : (X_1, \dots, X_n) \sim P_X, \qquad K : (X_1, \dots, X_n) \sim Q_X$$

where P_X and Q_X are complicated (not iid!) dist.

Bayesian resolution of "optimal test" problem

- Recall: Bayesianist has a prior for everything
- So he says: **Under null H, data is generated as:**
  1. Nature picks $\theta \in \mathcal{C}_0$ via $\theta \sim P^{(H)}$
     ($P_\theta^{(H)}$ is the secret sauce: the magic prior on θ's)
  2. Generate $X_i \overset{iid}{\sim} P^\theta$
- (similar under alt. $K : \theta \sim P^{(K)}$ etc.
- This is awesome: now we have a simple vs simple HT:

$$H : (X_1, \ldots, X_n) \sim P_X, \qquad K : (X_1, \ldots, X_n) \sim Q_X$$

  where P_X and Q_X are complicated (not iid!) dist.
- Neyman-Pearson: Likelihood-ratio is the **optimal test:**

$$\frac{P_X(X_1, \ldots, X_n)}{Q_X(X_1, \ldots, X_n)} \leq \gamma \quad \Rightarrow \quad \boxed{\text{REJECT}}$$

Bayesian intermission

Bayesian resolution of "optimal test" problem

- Recall: Bayesianist has a prior for everything
- So he says: **Under null H, data is generated as:**
  1. Nature picks $\theta \in \mathcal{C}_0$ via $\theta \sim P^{(H)}$
     ($P_\theta^{(H)}$ is the secret sauce: the magic prior on θ's)
  2. Generate $X_i \overset{iid}{\sim} P^\theta$
- (similar under alt. $K : \theta \sim P^{(K)}$ etc.
- This is awesome: now we have a simple vs simple HT:

$$H : (X_1, \ldots, X_n) \sim P_X, \qquad K : (X_1, \ldots, X_n) \sim Q_X$$

  where P_X and Q_X are complicated (not iid!) dist.
- Neyman-Pearson: Likelihood-ratio is the **optimal test:**

$$\frac{P_X(X_1, \ldots, X_n)}{Q_X(X_1, \ldots, X_n)} \leq \gamma \quad \Rightarrow \quad \boxed{\text{REJECT}}$$

- **Pros:** always have some test. **Cons:** hard to find good priors

Back to frequentist

Next topic: Examples of optimal tests for frequentists.

One-sided GLM test

- Setting:
  - $X \overset{iid}{\sim} \mathcal{N}(\theta, 1)$
  - known-variance, Gaussian location model

$$H : \theta = 0 \qquad \text{vs.} \qquad K : \theta > 0 \quad \text{(one-sided)}$$

  - Test is size-α iff $\beta(0) \leq \alpha$

One-sided GLM test

- Setting:
  - $X \overset{iid}{\sim} \mathcal{N}(\theta, 1)$
  - known-variance, Gaussian location model

$$H : \theta = 0 \qquad \text{vs.} \qquad K : \theta > 0 \quad \text{(one-sided)}$$

  - Test is size-α iff $\beta(0) \leq \alpha$
- Side problem: What is the best test for $\theta = 0$ vs. $\theta = a$?

One-sided GLM test

- Setting:
  - $X \overset{iid}{\sim} \mathcal{N}(\theta, 1)$
  - known-variance, Gaussian location model

$$H : \theta = 0 \qquad \text{vs.} \qquad K : \theta > 0 \quad \text{(one-sided)}$$

  - Test is size-α iff $\beta(0) \leq \alpha$
- Side problem: What is the best test for $\theta = 0$ vs. $\theta = a$?
- Neyman-Pearson's lemma: log-likelihood is optimal:

$$\begin{aligned}\tilde{T}_a(x_1, \ldots, x_n) &= -\sum_{i=1}^{n} \log \frac{P_{X|\theta=0}(x_i)}{P_{X|\theta=a}(x_i)} \\ &= (na)\bar{x}_n - na^2\end{aligned}$$

One-sided GLM test

- Setting:
  - $X \overset{iid}{\sim} \mathcal{N}(\theta, 1)$
  - known-variance, Gaussian location model

$$H : \theta = 0 \qquad \text{vs.} \qquad K : \theta > 0 \quad \text{(one-sided)}$$

  - Test is size-α iff $\beta(0) \leq \alpha$
- Side problem: What is the best test for $\theta = 0$ vs. $\theta = a$?
- Neyman-Pearson's lemma: log-likelihood is optimal:

$$\begin{aligned} \tilde{T}_a(x_1, \ldots, x_n) &= -\sum_{i=1}^{n} \log \frac{P_{X|\theta=0}(x_i)}{P_{X|\theta=a}(x_i)} \\ &= (na)\bar{x}_n - na^2 \end{aligned}$$

- Select threshold: $P[(na)\bar{X}_n - na^2 > t|\theta = 0] = \alpha$
- Final rule:

$$\sqrt{n}\bar{X}_n \geq \Phi^{-1}(1 - \alpha) \quad \Rightarrow \quad \boxed{\text{REJECT}}$$

where $\Phi^{-1}(q) = q$-th quantile of $\mathcal{N}(0, 1)$.

One-sided GLM test

- Setting:
  - $X \overset{iid}{\sim} \mathcal{N}(\theta, 1)$
  - known-variance, Gaussian location model

$$H : \theta = 0 \qquad \text{vs.} \qquad K : \theta > 0 \quad \text{(one-sided)}$$

  - Test is size-α iff $\beta(0) \le \alpha$
- Side problem: What is the best test for $\theta = 0$ vs. $\theta = a$?
- Neyman-Pearson's lemma: log-likelihood is optimal:

$$\tilde{T}_a(x_1, \ldots, x_n) \quad = \quad - \sum^{n} \log \frac{P_{X|\theta=0}(x_i)}{P_{X|\theta=a}(x_i)}$$

MAGIC: Rule does not depend on a

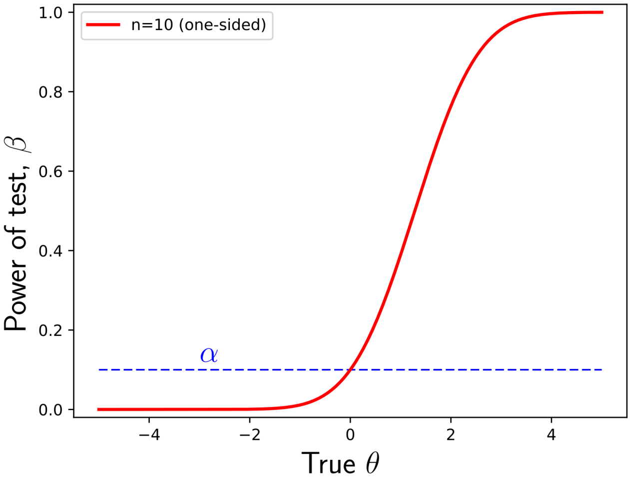
- Select threshold: $P[(na)\bar{X}_n - na^2 > t | \theta = 0] = \alpha$
- Final rule:

$$\sqrt{n}\bar{X}_n \ge \Phi^{-1}(1 - \alpha) \quad \Rightarrow \quad \text{REJECT}$$

where $\Phi^{-1}(q) = q$-th quantile of $\mathcal{N}(0, 1)$.

One-sided GLM test is UMP

$H : \theta = 0$ vs. $K : \theta > 0$



Punchline: one-sided z-test is unique most-powerful test.

Two-sided GLM test

- Setting:
  - $X \overset{iid}{\sim} \mathcal{N}(\theta, 1)$
  - known-variance, Gaussian location model

$$H : \theta = 0 \qquad \text{vs.} \qquad K : \theta \neq 0 \quad \text{(one-sided)}$$

  - Test is size-α iff $\beta(0) \leq \alpha$
- What is the best test for $\theta = 0$ vs. $\theta = a$?

Two-sided GLM test

- Setting:
  - $X \overset{iid}{\sim} \mathcal{N}(\theta, 1)$
  - known-variance, Gaussian location model

$$H : \theta = 0 \qquad \text{vs.} \qquad K : \theta \neq 0 \quad \text{(one-sided)}$$

  - Test is size-α iff $\beta(0) \leq \alpha$
- What is the best test for $\theta = 0$ vs. $\theta = a$?

$$\mathrm{sign}(a)\sqrt{n}\bar{X}_n \geq \Phi^{-1}(1-\alpha) \quad \Rightarrow \quad \boxed{\text{REJECT}}$$

where $\Phi^{-1}(q) = q$-th quantile of $\mathcal{N}(0, 1)$.

Two-sided GLM test

- Setting:
  - $X \overset{iid}{\sim} \mathcal{N}(\theta, 1)$
  - known-variance, Gaussian location model

$$H : \theta = 0 \qquad \text{vs.} \qquad K : \theta \neq 0 \quad \text{(one-sided)}$$

  - Test is size-α iff $\beta(0) \leq \alpha$
- What is the best test for $\theta = 0$ vs. $\theta = a$?

$$\mathrm{sign}(a)\sqrt{n}\bar{X}_n \geq \Phi^{-1}(1-\alpha) \quad \Rightarrow \quad \boxed{\text{REJECT}}$$

where $\Phi^{-1}(q) = q$-th quantile of $\mathcal{N}(0, 1)$.
- So there is no one rule to dominate them. No magic.
- What should we do?

Definition (Unbiased tests)

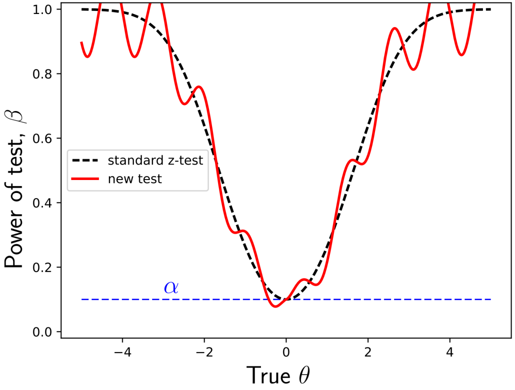
A size-α test T is unbiased if $\beta(\theta) \geq \alpha \quad \forall \theta \in K$

- **Important:** unbiasedness of T depends on both H and K
- Aka "test does not treat some $\theta \in K$ more favorably than $\theta \in H$"

Definition (Unbiased tests)

A size-α test T is unbiased if $\beta(\theta) \geq \alpha \quad \forall \theta \in K$

- **Important:** unbiasedness of T depends on both H and K
- Aka "test does not treat some $\theta \in K$ more favorably than $\theta \in H$"
- Examples: **black** – unbiased, **red** – biased.



Best unbiased tests (UMPU)

Definition (Unbiased tests)

A size-α test T is unbiased if $\beta(\theta) \geq \alpha \quad \forall \theta \in K$

- Setting:
  - $X \overset{iid}{\sim} \mathcal{N}(\theta, 1)$
  - known-variance, Gaussian location model **(GLM)**

$$H : \theta = 0 \qquad \text{vs.} \qquad K : \theta \neq 0 \quad \text{(one-sided)}$$

  - Test is size-α iff $\beta(0) \leq \alpha$
- We have seen, no most powerful test exist
- But there exists most powerful test among all unbiased.

Best unbiased tests (UMPU)

Definition (Unbiased tests)

A size-α test T is unbiased if $\beta(\theta) \geq \alpha \quad \forall \theta \in K$

- Setting:
  - $X \overset{iid}{\sim} \mathcal{N}(\theta, 1)$
  - known-variance, Gaussian location model **(GLM)**

$$H : \theta = 0 \qquad \text{vs.} \qquad K : \theta \neq 0 \quad \text{(one-sided)}$$

  - Test is size-α iff $\beta(0) \leq \alpha$
- We have seen, no most powerful test exist
- But there exists most powerful test among all unbiased.
- No surprise: it is our friend z-test
- ... it is unique such test! (aka **UMPU**).

Best unbiased tests (UMPU)

Definition (Unbiased tests)

A size-α test T is unbiased if $\beta(\theta) \geq \alpha \quad \forall \theta \in K$

Other examples of **UMPUs**:

- One-sample tests: $(X \overset{iid}{\sim} \mathcal{N}(\mu_X, \sigma_X^2))$
  - unknown-variance, one-sided: $\mu_X = 0$ vs $\mu_X > 0$
  - unknown-variance, two-sided: $\mu_X = 0$ vs $\mu_X \neq 0$

  UMPU: t-test (with threshold from Student-t)
- Two-sample tests: $(X, Y \overset{iid}{\sim} \mathcal{N})$
  - unknown variance, $\sigma_X = \sigma_Y$ two-sided: $\mu_X = \mu_Y$ vs $\mu_X \neq \mu_Y$
  - unknown variance, $\sigma_X = \sigma_Y$ one-sided: $\mu_X = \mu_Y$ vs $\mu_X > \mu_Y$

  UMPU: two-sample t-test (pooled variance)
  - unknown-variance, $\sigma_X \neq \sigma_Y$.

  UMPU: ??? (Behrens-Fisher problem)