Testing with one real parameter

Outline

1) Uniformly most powerful tests
2) Two-sided tests
3) p-Values
4) Confidence Regions
Uniformly most powerful tests

General setup: \( \mathcal{H}, \Theta_0, \Theta_1 \)

**Def.** If \( \phi^*(x) \) has sig. level \( \alpha \), and for any other level-\( \alpha \) test \( \phi \) we have
\[
E_{\Theta_0} \phi^* \geq E_{\Theta_0} \phi \quad \forall \Theta_0 \in \Theta
\]
then \( \phi^* \) is **uniformly most powerful (UMP)**.

Typically only exist for 1-sided testing in certain 1-parameter families.

**Def.** A model \( \mathcal{P} \) is **identifiable** if
\[
\Theta_1 \neq \Theta_2 \Rightarrow \mathbb{P}_{\Theta_1} \neq \mathbb{P}_{\Theta_2} \quad \text{(exists: } \mathbb{P}_{\Theta_1}(A) \neq \mathbb{P}_{\Theta_2}(A))
\]

**Def.** Assume \( \mathcal{P} = \{ \mathbb{P}_\Theta : \Theta \in \Theta \subseteq \mathbb{R}^3 \} \) has densities \( \mathbb{P}_\Theta \), and is identifiable. We say \( \mathcal{P} \) has
**monotone likelihood ratios (MLR)** if
there is some statistic \( T(X) \) s.t.
\[
\frac{\mathbb{P}_{\Theta_2}}{\mathbb{P}_{\Theta_1}}(X) \text{ is a nondecreasing function of } T(X),
\]
for \( \Theta_1 < \Theta_2 \) \( [ \text{some } T(X) \text{ for all } \Theta \text{'s} ] \)
\[
\left( \frac{\frac{c}{o}}{c} = \infty \text{ if } c > 0, \quad \frac{d}{o} \text{ undefined} \right)
\]

**Ex.** Exp. fam: \( e^{(\gamma_1 - \gamma_0)T(X) - n(A(\gamma_1) - A(\gamma_0))} \uparrow \) in \( \mathcal{E}(X) \)
Theorem: Assume \( \mathcal{Y} \) has MLE test \( H_0: \theta \leq \theta_0 \) vs \( H_1: \theta > \theta_0 \) at level \( \alpha = (0, 1) \)

Let \( \phi^*(x) = \begin{cases} 0 & T(x) < c \\ \gamma & T(x) = c \\ 1 & T(x) > c \end{cases} \)

with \( c, \gamma \) chosen so \( E_{\theta_0} \phi^*(x) = \alpha \in (0, 1) \)

a) \( \phi^* \) is a UMP level-\( \alpha \) test

b) \( \beta(\theta) = E_{\theta} \phi^*(x) \) is non-decreasing in \( \theta \), strictly incr. wherever \( \beta(\theta) \in (0, 1) \)

c) If \( \theta_1 < \theta_0 \), then \( \phi^* \) minimizes \( E_{\theta_0} \phi(x) \)

among all tests with \( E_{\theta_0} \phi(x) = \alpha \)
Proof
b) Suppose $\Theta_1 < \Theta_2$, then \( \frac{p_{\Theta_2}}{p_{\Theta_1}}(x) \) is a non-decreasing function of \( T(x) \)

\[ \Rightarrow \phi^* \] is a LRT for \( H_0: \theta = \Theta_1 \) vs. \( H_1: \theta = \Theta_2 \) at level \( \alpha = E_{\Theta_1} \phi^*(x) \)

By Cor. 12.4, \( E_{\Theta_2} \phi(x) \geq E_{\Theta_1} \phi(x) \), strict ineq. unless both \( = 0 \) or \( 1 \)

a) Suppose \( \Theta_1 > \Theta_0 \) and \( \phi \) has level \( \leq \alpha \)

\[ \Rightarrow E_{\Theta_1} \phi^*(x) = E_{\Theta_0} \phi(X) \] since \( \phi^* \) is a LRT for \( H_0: \theta = \Theta_0 \) vs. \( H_1: \theta = \Theta_1 \)

c) \( \Theta_1 < \Theta_0 \), assume \( E_{\Theta_0} \tilde{\phi}(x) = E_{\Theta_0} \phi^*(x) = \alpha \)

Both \( 1-\tilde{\phi} \), \( 1-\phi^* \) are tests of \( H_0: \theta = \Theta_0 \) vs. \( H_1: \theta = \Theta_1 \) both have sig. level \( 1-\alpha \)

\[ 1-\phi^* \] is a LRT since \( \frac{p_{\Theta_1}}{p_{\Theta_0}}(x) \) is non-incr. in \( T(x) \)

\[ \Rightarrow E_{\Theta_0}(1-\tilde{\phi}) \leq E_{\Theta_0}(1-\phi^*) = 1-\alpha \]

Intuition \( \phi^* \) is a LRT for \( H_0: \Theta = \Theta_0 \) vs. \( H_1: \Theta = \Theta_1 \) for any pair \( \Theta_0 < \Theta_1 \) (sig. level depends on \( \Theta_0 \))

[This lets us extend our simple vs. simple result to (a very special case of) composite vs comp.]
Two-sided Alternatives

Setup: \( \mathcal{P} = \{ \mathcal{P}_\theta : \theta \in \Theta \subseteq \mathbb{R} \} \), \( \Theta_0 \in \Theta \)

Test \( H_0 : \theta = \theta_0 \) vs. \( H_1 : \theta \neq \theta_0 \)

(Can be generalized naturally to \( H_0 : \theta \in [\theta_1, \theta_2] \))

Def: A real-valued statistic \( T(X) \) is stochastically increasing in \( \theta \) if

\[ \mathbb{P}_\theta (T(X) \leq t) \text{ is non-incr. in } \theta, \forall t \]

Assume \( T(X) \) is a stochastically increasing summary test statistic

\[ \begin{align*} \text{Ex: } & X_i \overset{iid}{\sim} \rho(x - \theta) \text{ (location family)} \\ & T(X) = \text{sample near/median} \\ \text{Ex: } & X_i \overset{iid}{\sim} \frac{1}{\theta} \rho(x/\theta) \text{ (scale family)} \\ & T(X) = \sum X_i^2 \text{ or median}(|X_1|, \ldots, |X_n|) \end{align*} \]
Two-tailed test rejects when $T(X)$ is "extreme": 

$$
\phi(x) = \begin{cases} 
1 & T(X) > c_2 \text{ or } T(X) < c_1 \\
0 & T(X) \in (c_1, c_2) \\
\gamma_i & T(X) = c_i 
\end{cases}
$$

Let $\alpha_1 = P_{\Theta_0}(T < c_1) + \gamma_1 P_{\Theta_0}(T = c_1)$ 

$\alpha_2$ similar for upper tail

Need $\alpha_1 + \alpha_2 = \alpha$, but how to balance?

Idea 1: Equal-tailed test: $\alpha_1 = \alpha_2 = \frac{\alpha}{2}$

Def: $\phi(x)$ is unbiased if $\inf_{\theta \in \Theta} E_\theta \phi(x) \geq \alpha$

Idea 2: Unbiased test: ensure $\min_{\theta} \beta_\phi(\theta) = \alpha$

(usually $\iff \frac{d\beta_\phi(\theta)}{d\theta} = 0$)
Theorem Assume $X_i \overset{iid}{\sim} e^{\Theta T(x) - A(x)} h(x)$

$H_0: \Theta \in [\theta_1, \theta_2]$ vs $H_1: \Theta < \theta_1$ or $\Theta > \theta_2$

(possibly $\theta_1 = \theta_2$)

Then

a) The unbiased test based on

$\sum T(X_i)$ with sig. level = $\alpha$ is UMP among all unbiased tests (UMPU)

b) If $\theta_1 < \theta_2$ the UMPU test can be found by solving for $c_i, \gamma_i$

\[ s.t. \quad E_{\Theta_1} \phi = E_{\Theta_2} \phi = \alpha \]

c) If $\theta_1 = \theta_2 = \theta_0$ the UMPU test can be found by solving for $c_i, \gamma_i$

\[ s.t. \quad E_{\Theta_0} \phi(x) = \alpha \quad \text{and} \quad \frac{d}{d\theta} \phi(\theta_0) = \left[ E_{\Theta_0} \left[ \sum T(X_i) \phi(x) - \alpha \right] \right] = 0 \]

(Proof in Keener)
**p-Values**

**Informal definition:** Suppose \( p(x) \) rejects for large values of \( T(x) \).

\[
p(x) = \Pr_{H_0} \left( T(X) \geq T(x) \right)
= \sup_{\Theta \in \Theta_0} \Pr_{\Theta} \left( T(X) \geq T(x) \right)
\]

**Example**

\( X \sim N(0,1) \) \quad H_0 : \Theta = 0 \quad \text{vs.} \quad H_1 : \Theta \neq 0

Two-sided test rejects for large \( T(X) = |X| \)

\[
\Leftrightarrow p(x) = 1 \{ |X| > z_{\alpha/2} \}
\]

The two-sided \( p \)-value is \( p(x) \) where

\[
p(x) = \Pr_{\Theta} \left( |X| > |x| \right)
= 2 \left( 1 - \Phi(|x|) \right)
\]

For \( H_0 : |\Theta| < \delta \quad \text{vs.} \quad H_1 : |\Theta| > \delta \):

\[
p(x) = \Pr_{\delta} \left( |X| > |x| \right)
= \Pr_{\delta} \left( \max \left( |X| - \delta, 0 \right) \right)
= 1 - \Phi \left( |x| - \delta \right) + \Phi \left( -|x| - \delta \right)
\]

etc.
Formal definition: \( \Theta, \Omega_0, \Theta_0 \).

Assume we have a test \( \phi_\alpha \) for each significance level, \( \sup_{\Theta \in \Theta_0} E_{\Theta} \phi_\alpha(X) \leq \alpha \)  

(non-randomized case: \( \phi_\alpha = 1 \{ x \in R_\alpha \} \))

Assume tests are monotone in \( \alpha \):  
if \( \alpha_1 \leq \alpha_2 \) then \( \phi_{\alpha_1}(x) \leq \phi_{\alpha_2}(x) \)  
(non-randomized: \( R_{\alpha_1} \leq R_{\alpha_2} \))

Then \( \rho(x) = \inf \{ \alpha : \phi_{\alpha}(x) = 1 \} \)  
(= \( \inf \{ \alpha : x \in R_{\alpha} \} \))
(possible to define randomized \( \rho \)-value but not worth it)

Note \( \rho(x) \leq \alpha \iff \phi_{\rho(x)}(x) = 1 \ \forall \ \alpha > \rho(x) \)

For \( \Theta \in \Omega_0 \), \( \mathbb{P}_\Theta(\rho(x) \leq \alpha) \leq \inf_{\alpha > \rho(x)} \mathbb{P}_\Theta(\phi_{\alpha}(X) = 1) \leq \alpha \)

\( \Rightarrow \) \( \rho \)-value stochastically dominates \( u[0,1] \)

If \( \phi_\alpha \) rejects for large \( \Gamma(X) \), coincides with informal definition
Note the p-value depends on
- the model & null hyp.,
- the data, AND
- the choice of test

\[ X \sim N_d(\Theta, I_d) \]
\[ H_0 : \Theta = 0 \text{ vs } H_1 : \Theta \neq 0 \]

We can use
\[ T_1(x) = \| X \|^2 \] (\( \chi^2 \) test)
\[ T_2(x) = \| X \|_\infty \] (max test)
\[ = \max_i |X_{i,:}| \]

Very different p-values / power if \( d \) large
(choice reflects belief about whether \( \Theta \) is sparse)
Confidence Sets

Accept/reject decision only so interesting:

- usually we care how big $\Theta$ is
- tiny $p$-value doesn't imply big $\Theta$

(big $p$-value doesn't imply small $\Theta$ either)

**Def.** $P = \{ \Theta : \Theta \in \Theta \}$

$C(X)$ is a **confidence set** for $g(\Theta)$ if

$$P_\Theta (C(X) \ni g(\Theta)) \geq 1 - \alpha, \quad \forall \Theta \in \Theta$$

We say $C(X)$ **covers** $g(\Theta)$ if $C(X) \ni g(\Theta)$

$P_\Theta (C(X) \ni g(\Theta))$ is **coverage probability**

$$\inf_\Theta P_\Theta (C \ni g(\Theta))$$ is **conf. level**

**Notes**

- $C(X)$ is random, not $g(\Theta)$
- Often misinterpreted as Bayesian guarantee

Say "$C(X)$ has a 95% chance of covering"

**NOT** "$g(\Theta)$ has a 95% chance of being in $C"

NEVER "95% chance $g(\Theta) \in [0.5, 1.5]" (e.g.)
Suppose we have a level-\(\alpha\) test \(\phi(x; a)\) of \(H_0: g(\theta) = a\) vs. \(H_1: g(\theta) \neq a, \forall a \in g(\Theta)\).

We can use it to make a confidence set for \(g(\theta)\):

Let \(\mathcal{C}(x) = \{a: \phi(x; a) < 1\}\)

= "all non-rejected values of \(\theta\)"

Then \(P_{\theta}(\mathcal{C}(x) \neq g(\theta)) = P_{\theta}(\phi(x; g(\theta)) = 1) \leq \alpha \quad \forall \theta\)

Alternatively, suppose \(\mathcal{C}(x)\) is a 1-\(\alpha\) confidence set for \(g(\theta)\).

We can use \(\mathcal{C}\) to construct a test \(\phi(x)\) of

\(H_0: g(\theta) = a\) vs. \(H_1: g(\theta) \neq a\)

\(\phi(x) = 1\{a \notin \mathcal{C}(x)\}\)

For \(\theta\) s.t. \(g(\theta) = a:\)

\(\mathbb{E}_{\theta} \phi(x) = P_{\theta}(\mathcal{C}(x) \neq g(\theta)) \leq \alpha\)

This is called inverting the test.
Confidence Intervals / Bounds

If \( C(X) = [C_1(x), C_2(x)] \) we say \( C(X) \) is a confidence interval (CI)

\[ C(X) = [C_1(x), \infty) : \text{lower conf. bd. (LCB)} \]
\[ C(X) = (-\infty, C_2(x)] : \text{upper conf. bd. (UCB)} \]

We usually get LCB / UCB by inverting a one-sided test in appropriate direction.
Called uniformly most accurate (UMA) if test UMP.

Get CI by inverting a two-sided test.
Called UMAU if test is UMPU.
\[ X \sim \text{Exp}(\theta) = \frac{1}{\theta} e^{-x/\theta}, \quad x > 0, \quad \theta > 0 \]

CDF \quad P_\theta(X \leq x) = 1 - e^{-x/\theta}

\underline{LCB}: Invert test for \( H_0: \theta \leq \theta_0 \)

\[ \text{Solve } \alpha = \int_{c(\theta_0)}^{\infty} = e^{-c(\theta_0)/\theta_0} \]

\[ c(\theta_0) = -\theta_0 \log \alpha \quad (\geq 0) \]

\[ X \leq c(\theta_0) \Rightarrow \theta_0 \geq \frac{X}{-\log \alpha} \]

\[ C(X) = \left[ \frac{X}{-\log \alpha}, \infty \right) \]

\underline{UCB}: Similar, \( C(X) = (-\infty, \frac{X}{-\log(1-\alpha)}) \]

\underline{Equal-tailed CI}:

Invert equal-tailed test of \( H_0: \theta = \theta_0 \)

\[ \phi_{\alpha/2}^2(X) = \phi_{-\theta_0}^{+\theta_0}(X) + \phi_{-\theta_0}^{-\theta_0}(X) \]

\[ H: \theta = \theta_0 \quad \text{H}_0: \theta \geq \theta_0 \quad \text{H}_0: \theta \leq \theta_0 \]

\[ \Rightarrow C(X) = \left[ \frac{X}{-\log \alpha}, \infty \right) \cap (-\infty, \frac{X}{-\log(1-\alpha)}) \]

\[ = \left[ \frac{X}{-\log \alpha}, \frac{X}{-\log(1-\alpha)} \right] \]
(Mis-)Interpreting Hypothesis Tests

Hypothesis tests ubiquitous in science

Common misinterpretations:

1) \( p < 0.05 \) therefore "there is an effect"
   or "the effect size = the estimate"

2) \( p > 0.05 \) therefore "there is no effect"

3) \( p = 10^{-6} \) therefore "the effect is huge"

4) \( p = 10^{-6} \) therefore "the data are significant"
   and everything about our model is correct in most naive interp.

5) Effect CI for men is \([0.2, 3.2]\),
   for women is \([-0.2, 2.8]\) therefore
   "there is an effect for men and not for women."

378) We rejected a specific, parametric null model therefore something interesting is happening.
How to interpret testing

Learning about the world from data is not easy or automatic!

Hypothesis tests let us ask specific questions about specific data sets under specific modeling assumptions, using specific testing methods.

All of these choices bear on the interpretation.

Top-tier medical journals let people publish claims, reporting p-values without saying what model was used or what test was employed.

**THIS IS ABSOLUTELY OUTRAGEOUS**

Hyp. tests can be a good companion to critical thinking, never a substitute.

“All models are wrong, some are useful” but need experience and theory to understand when assumptions do or don’t cause real trouble.
Conceptual Objections

Q1: Why should I test \( H_0 : \theta = 0 \)? No \( \theta \) is ever exactly 0.

A1: a) Test \( H_0 : |\theta| \leq \delta \) if you want!
   If \( \text{s.e.}(\hat{\theta}) = 10\delta \), not much difference.
   b) Most two-sided tests justify directional inference:
   "If \( T > c \), declare \( \theta > 0 \), if \( T < c \),
   declare \( \theta < 0 \)," with \( \text{Pr} \text{(false claim)} \leq \alpha \)
   c) Harder to answer in non-parametric problems,
   e.g. \( H_0 : P = Q \) vs \( H_1 : P \neq Q \) for
   perm. test, but alternative frameworks like
   Bayes force very strong assumptions on us.

Q2: People only like frequentist results like
    \( p \)-values, \( \text{CI's} \) because they mistake them
    for Bayesian results.
    95% chance \( C(X) \ni \theta \) is misinterpreted as
    a claim about \( \text{Pr}(\theta | X) \).

A2: True, but subjective Bayesian results often
    misinterpreted as "the posterior dist. of \( \theta \"
    when really should be "my posterior opinion about \( \theta \)"
b) “Objective” Bayesian credible intervals are even worse: “nobody’s posterior opinion about θ”

c) **Caveat**: in some simple, low-dim., high signal settings, can maybe say “any reasonable person’s posterior opinion about θ.” Then Bayes methods probably best!

Q3: $p$-values ignore $P(\text{Data} \mid H_1)$ and only look at $P(\text{Data} \mid H_0)$. Data might be more likely under $H_0$ but still reject.

A3: $P(\text{Data} \mid H_0)$, $P(\text{Data} \mid H_1)$ only make sense for simple null/alternative. Even in $N(0, 1)$ $H_0: \theta = 0$, what is $P(X = 1 \mid H_1)$?

If $H_1$ is vague prior like $\Theta \sim N(0, 10^6)$, then $X \sim N(0, 10^6 + 1)$ and $P(4 \mid H_1) \ll P(4 \mid H_0)$ Will scientists understand this??

Even bigger problems in high-dim., hierarchical, or nonparametric priors.
Q4: Scientists always misuse hypothesis testing, so we should switch to something else
(Confidence intervals / Bayes / weird new idea)

A4: a) CIs great, but just a re-packaging of hypothesis tests. (Might still be good for staving off some common misinterpretations by naifs)

b) Bayes has its uses, but forcing scientists to make more choices/assumptions is not going to solve problem of scientist incentives/ignorance

c) Most weird new ideas have bigger issues but just haven’t been criticized much yet b/c no one but proponents care.

d) Statistical inference will never be idiot-proof, b/c science/critical thinking are not idiot-proof. Engineers have to learn calculus, learning what a p-value means is not that hard. Suck it up, social scientists!
(and ask for help)