Outline

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

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

**Def.** If \( \phi^*(x) \) has sig. level \( \alpha \), and for any other level-\( \alpha \) test \( \phi \) we have

\[
E_{\Theta} \phi^* \geq E_{\Theta} \phi \quad \forall \Theta \in \Theta_1, \\
\text{then } \phi^* \text{ is uniformly most powerful (UMP)}
\]

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

**Def.** A model \( \mathcal{P} \) is identifiable if

\[
\Theta_1 \neq \Theta_2 \implies P_{\Theta_1} \neq P_{\Theta_2} \quad (E A: P_{\Theta_1}(A) \neq P_{\Theta_2}(A))
\]

**Def.** Assume \( \mathcal{P} = \{P_\Theta: \Theta \in \Theta \subseteq \mathbb{R}^3 \} \) has densities \( p_\Theta \), and is identifiable. We say \( \mathcal{P} \) has monotone likelihood ratios (MLR) if

there is some statistic \( T(X) \) s.t.

\[
\frac{p_{\Theta_2}(X)}{p_{\Theta_1}(X)} \text{ is a nondecreasing function of } T(X),
\]

for any \( \Theta_1 < \Theta_2 \) \( [ \text{some } T(X) \text{ for all } \Theta \text{'s} ] \)

\[
(\frac{c_0}{c_0} = \infty \text{ if } c > 0, \quad \frac{0}{0} \text{ undefined})
\]

**Ex.** Exp. fam: \( e^{(x, -\gamma_0) \cdot T(x) - v(A(\gamma_1) - A(\gamma_0))} \searrow \text{ in } \Sigma T(X) \)
Theorem \ Assume \ \mathcal{P} \ has \ MLR, \ 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 \ \ \mathbb{E}_{\theta_0} \phi^*(x) = \alpha \in (0, 1) 

a) \ \phi^* \ is \ a \ UMP \ level-\alpha \ test \\
b) \ \beta(\theta) = \mathbb{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 \ \mathbb{E}_{\theta} \phi(x) 
among \ all \ tests \ with \ \mathbb{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 $\leq \alpha$

$\Rightarrow E_{\Theta_1} \phi^*(X) = E_{\Theta_1} \tilde{\phi}(X)$ 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 $1-\phi^*$, $1-\tilde{\phi}$ 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 $\frac{P_{\Theta_1}}{P_{\Theta_2}}(X)$ is non-incr. in $T(X)$

$\Rightarrow E_{\Theta_0}(1-\tilde{\phi}) \leq E_{\Theta_1}(1-\phi^*) = 1-\alpha \Box$

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: \[ P = \{ \Theta_0 : \Theta \in \Theta \subseteq \mathbb{R}^n, \Theta_0 \in \Theta \} \]

Test \[ H_0 : \Theta = \Theta_0 \] vs. \[ H_1 : \Theta \neq \Theta_0 \]

(Can be generalized naturally to \[ H_0 : \Theta \in [0, \Theta_2] \])

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

\[ \Pr_{\Theta}(T(X) \leq t) \] is non-increas. in \( \Theta, \forall t \)

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

**Ex** \( X_i \overset{iid}{\sim} \rho(x - \Theta) \) (location family)

\[ T(X) = \text{sample near/median} \]

**Ex** \( X_i \overset{iid}{\sim} \frac{1}{\Theta} \rho(x/\Theta) \) (scale family)

\[ T(X) = \sum X_i^2 \text{ or median}(|X_1|, \ldots, |X_n|) \]
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 $\alpha_1 = \Pr_{\Theta_0}(T < c_1) + \gamma_1 \Pr_{\Theta_0}(T = c_1)$

$\alpha_2$ 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 $\inf_{\Theta} \mathbb{E}_\Theta \phi(x) \geq \alpha$

**Idea 2:** Unbiased test: ensure $\min \beta_\phi(\Theta) = \alpha$

(usually $\iff \frac{d\beta_\phi(\Theta)}{d\Theta}(\Theta_0) = 0$)

Def $\phi(x)$ is unbiased if $\inf_{\Theta} \mathbb{E}_\Theta \phi(x) \geq \alpha$

**Idea 2:** Unbiased test: ensure $\min \beta_\phi(\Theta) = \alpha$

(usually $\iff \frac{d\beta_\phi(\Theta)}{d\Theta}(\Theta_0) = 0$)
Theorem. Assume $X_i \sim \theta T(x) - A(x)$, $n(x)$

$H_0: \theta \in [\theta_1, \theta_2]$ 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 (UMP-U).

b) If $\theta_1 < \theta_2$ the UMP-U test can be found by solving for $c_i, x_i$
s. t. $E_{\theta_1} \phi = E_{\theta_2} \phi = \alpha$

c) If $\theta_1 = \theta_2 = \theta_0$ the UMP-U test can be found by solving for $c_i, x_i$
s. t. $E_{\theta_0} \phi(x) = \alpha$ and

$$\frac{d\beta_0(\theta_0)}{d\theta} = E_{\theta_0} \left[ \sum T(x_i)(\phi(x_i) - \alpha) \right] = 0$$

(Proof in Keener)
\textbf{p-Values}

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

\[
p(x) = \sup_{\theta \in \Theta_0} P_\theta (T(X) \geq T(x)) \]

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

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

\[
(\Leftrightarrow \phi(x) = 1 \{ |X| > \frac{\sigma}{\sqrt{2}} \})
\]

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

\[
p(x) = \sup_{\theta \in \Theta_0} P_\theta (|X| > 1|X|) \]

\[
= 2 \left( 1 - \Phi(|X|) \right)
\]

For \( H_0: |\theta| < \delta \) vs. \( H_1: |\theta| > \delta \):

\[
p(x) = \sup_{\delta} P_\delta (|X| > 1|X|) \quad (= \sup_{\delta} (\text{"}\cdot\text{"}))
\]

\[
= 1 - \Phi(|X| - \delta) + \Phi(-|X| - \delta)
\]

etc.
**Formal definition:** \( \exists \), \( \Theta_0 \), \( \Theta \).

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

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

Assume tests are monotone in \( \alpha \):

if \( \alpha_1 \leq \alpha_2 \) then \( \phi_{\alpha_1}(x) \leq \phi_{\alpha_2}(x) \)

(non-randomized: \( R_{\alpha_1} \subseteq R_{\alpha_2} \))

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 \( \rho(x) \leq \alpha \Leftrightarrow \phi_{\tilde{\alpha}}(x) = 1 \ \forall \ \tilde{\alpha} > \alpha \)

For \( \Theta \in \Theta_0 \), \( \bar{\rho}_\Theta(\rho(x) \leq \alpha) \leq \inf_{\tilde{\alpha} > \alpha} \bar{\rho}_\Theta(\phi_{\tilde{\alpha}}(x) = 1) \leq \alpha \)

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

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

$\begin{align*}
  \text{Ex} & \quad \mathbf{x} \sim N_d(\Theta, \mathbf{I}_d) \quad H_0 : \Theta = 0 \; \text{vs} \; H_1 : \Theta \neq 0 \\
  \text{We can use} & \quad T_1(x) = \|x\|^2 \; (\chi^2 \text{ test}) \\
  \quad \text{or} & \quad T_2(x) = \|x\|_\infty \; (\text{max test}) \\
                        & \quad = \max_i |x_i|
\end{align*}$

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)]

\[ \text{Def } \ P = \{ \theta : \theta \in \Theta \} \]

\( C(X) \) 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 \( C(X) \ni g(\theta) \)

\[ P_\theta ( C(X) \ni g(\theta) ) \text{ is coverage probability} \]

\[ \inf_\theta P_\theta ( C \ni g(\theta) ) \text{ is conf. level} \]

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.)
Duality of Testing & Confidence Sets

Suppose we have a level-\(\alpha\) test \(\phi(x; \theta)\) of \(H_0: g(\theta) = a\) vs. \(H_1: g(\theta) \neq a, \forall \theta \in g(\theta)\).

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

\[ C(x) = \{ \theta : \phi(x; \theta) < 1 \} \]

be "all non-rejected values of \(\Theta\)."

Then

\[ P_\theta \left( C(x) \neq g(\theta) \right) = P_\theta \left( \phi(x; g(\theta)) = 1 \right) \leq \alpha \quad \forall \Theta \]

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

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

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

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

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

\[ \mathbb{E}_\theta \phi(x) = P_\theta \left( C(x) \neq g(\theta) \right) \leq \alpha \]

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 is UMP

Get CI by inverting a two-sided test

Called UMAU if test is UMPU
\[ X \sim \text{Exp}(\theta) = \frac{1}{\theta} e^{-x/\theta} \quad x > 0, \theta > 0 \]

CDF: \[ \mathbb{P}_\theta(X \leq x) = 1 - e^{-x/\theta} \]

**LCB**: Invert test for \( H_0 : \theta = \theta_0 \)

\[ \text{Solve} \quad \alpha = \mathbb{P}_\theta(X > c(\theta_0)) = e^{-c(\theta_0)/\theta_0} \]

\[ c(\theta_0) = \theta_0 \log(1/\alpha) \quad (> 0) \]

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

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

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

**Equal-tailed CI**:

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

\[ \Phi^{2T}_\alpha (X) = \Phi^{2\theta_0}_\alpha (X) + \Phi^{\theta_0}_\alpha (X) \]

2-tailed \( H_0 : \theta = \theta_0 \)

\[ H_0 : \theta \geq \theta_0 \quad H_0 : \theta \leq \theta_0 \]

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

\[ = \left[ \frac{X}{-\log \alpha} , \frac{X}{-\log(1-\alpha)} \right] \]
(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 interpretation.

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.

Pretty outrageous when you think about it!

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 $T > c_a$, declare $\theta > 0$, if $T < c$, declare $\theta < 0$" with $P(\text{false claim}) \leq \alpha$

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, CIs because they mistake them for Bayesian results.

95% chance $C(X) \ni \theta$ is misinterpreted as a claim about $p(\theta | X)$.

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 \( P(\text{Data} \mid H_i) \) and only look at \( P(\text{Data} \mid H_0) \). Data might be more likely under \( H_0 \) but still reject.

A3: \( P(\text{Data} \mid H_0) \), \( P(\text{Data} \mid H_i) \) only make sense for simple null/alternative. Even in \( N(\theta, 1) \), \( H_0: \theta = 0 \), what is \( P(X = 2.2 \mid H_i) \)?

If \( H_i \) is vague prior like \( \Theta \sim N(0, 10^6) \), then \( X \sim N(0, 10^6 + 1) \) and \( P(4 \mid H_i) < P(4 \mid H_0) \).

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)

(my opinions)

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. Ask for help if you need it!

(Corollary for Statisticians: don't be jerks)