Lecture 13
Hypothesis Testing

Goal: To assess the validity of certain claims or answer questions about a population based on data.

- Two claims
- Which of the two claims do the data support? (i.e. which of the claims is more consistent with the observed data?)

Qu: Is treatment A effective at treating Covid?

- " " " more effective than treatment B?
- Consider modeling the score on final exam as a linear regression as:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \cdots + \beta_m X_m + \text{error}$$

Y: score on the final; X_1: score on MT1; X_2: score on MT2; X_3: score on HWs etc

Qu: Is variable X_k helpful in predicting Y?

Qu: Is the coin fair? (i.e. a) $p = \text{Prob}(H) = 0.5$ or (ii) $p \neq 0.5$)

If (i) is true, then X_i = outcome on flip $i \sim \text{Bernoulli}(0.5)$
& for iid X_i, $Y_n = \sum_{i=1}^{n} X_i \sim \text{Binomial}(n, 0.5)$

Ex:

(a) $D = \{4H, 6T\}$

(b) $D = \{40H, 60T\}$

(c) $D = \{400H, 600T\}$

Which provide strongest evidence for (ii) " " " " against (i)?

A HT helps us assess validity of a claim or assertion (i.e. hypothesis) based on a data set (taking sampling variability & distr. into account).

- A HT consists of two hypotheses (non-overlapping)
  H_0: null hypothesis, status quo, usually requires no change
  vs
  H_a: alternative hypothesis, the claim to be assessed compared to H_0, usually requires substantial action.

- A HT is asymmetric in deciding the validity of H_0 vs H_a.
  - If there is sufficient evidence for H_a in data, you reject H_0 and accept H_a.
  - If there is no sufficient evidence for H_a in data, you fail to reject H_0 (or retain H_0).
  - One assumes population is as in H_0, and tests this assumption with evidence in data. If data (evidence) contradicts "substantially" with H_0, then s/he rejects H_0 in favor of H_a.

Ex: Coin fair or not! $H_0: p = 0.5$ vs $H_a: p \neq 0.5$

$D = \{4H, 6T\}$ reject or retain H_0?

Some examples of Ho & Ha:

Situation	Null Hypothesis	Alternative Hyp.
• Effect of a treatment on a disease	not effective (i.e. like a placebo)	effective
• Medical testing (for a disease)	Healthy / not infected	infected
• Drug test, A vs B	no difference	difference (eg. A better than B)
• extraterrestrial life in exoplanets	none (ie. no Earth-like life)	Earth life / life exists
* US Criminal System	innocent (not guilty)	guilty
• medical mask vs cloth "	no difference	medical mask better

~~etc.~~

Criminal Trial Analogy:

Ho: not guilty

Ha: guilty

• need evidence beyond reasonable doubt to declare the suspect guilty.

If no such evidence exists, suspect is not guilty (ie. evidence did not corroborate the guilt verdict)

Outline of a generic HT:

0. Set up H_0 & H_1 (if not provided)

1. Determine a statistic (as a measure of evidence from data) $T = T(D) = T(X_1, \dots, X_n)$ and compute it.

2. Determine a region of rejection (i.e. rejection region (RR)) for when or which values of T constitute "enough" evidence for H_a.

3. Check if T falls in RR or not.
   If T falls in RR, reject H_0
   " " does not fall in RR, fail to reject H_0 or retain H_0.

In general Hypotheses are written in terms of parameter

Ex: $X_1, \dots, X_n$ iid $f(x;\theta)$, θ is unknown parameter

simple null → $H_0: \theta = \theta_0$ vs $H_a: \theta \neq \theta_0$ (2-sided alter.)

one-sided tests:
- $H_0: \theta \leq \theta_0$ vs $H_a: \theta > \theta_0$ (right-sided alternative)
- or $H_0: \theta \geq \theta_0$ vs $H_a: \theta < \theta_0$ (left-sided alternative)

Ex: (Coin example) $H_0: p = 0.5$ vs $H_a: p \neq 0.5$ (2-sided)

if $H_0: p \leq 0.5$ vs $H_a: p > 0.5$ (1-sided)

- In a HT, one can make two types of error:

Truth \ Decision	Retain H_0	Reject H_0
H_0	✓ true -ive	X Type I error, false +ve
H_a	X Type II error, false -ve	✓ true +ve

Type I error: Rejecting a true H_0

" II " : Retaining a false H_0 (when H_a is true)

$P(\text{Type I error}) = \alpha$ = level of significance

$P(\text{Type II error}) = \beta$

$P(\text{reject } H_0 \mid H_0 \text{ is false}) = \text{power} = 1-\beta$.

- Want α & β small i.e. want small α & high power
- may not be possible to attain both goals, instead one controls $P(\text{type I error}) \leq \alpha$ and aims for high(est) power.

Types of Tests

1) Wald's Test: (aka Normal-Based Test) $D = \{X_1, \ldots, X_n\}$ $X_i \overset{iid}{\sim} F(x;\theta)$, θ unknown

$H_0: \theta = \theta_0$ vs $H_a: \theta \neq \theta_0$

$\hat{\theta}$ is an estimator for θ

$\hat{\theta}$ is asymptotically normal, i.e. $\hat{\theta} \overset{\text{approx}}{\sim} N(\theta, se^2)$

Assume H_0 is true, then $\hat{\theta} \overset{\text{approx}}{\sim} N(\theta_0, se^2)$
(i.e. under H_0, $\hat{\theta} \overset{\text{approx}}{\sim} N(\theta_0, se^2)$).

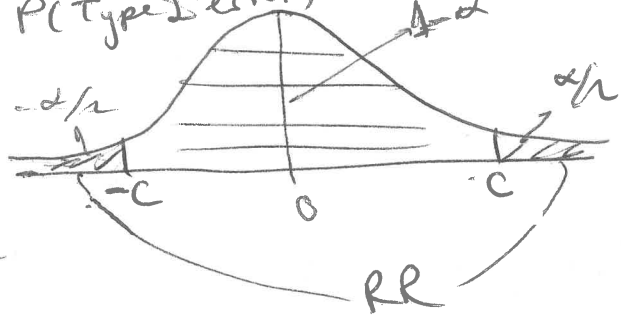
Statistic: $\hat{\theta}$ or better than this

$$W = \frac{\hat{\theta} - \theta_0}{se} \overset{\text{approx}}{\sim} N(0,1)$$

← if not known, replace with a consistent estimator $\hat{se}$.

Decision Rule (DR): Reject H_0 if $|\hat{\theta} - \theta_0|$ is large; i.e.
" " " $|W| > c$
(that is RR: $\{w : w > c$ or $w < -c\}$

But to have $P(\text{Type I error}) = \alpha$,
we need to have $P(W > c \mid H_0) + P(W < -c \mid H_0) = \alpha$



Hence, $P(W > c) = P(W < -c) = \alpha/2$

$\Rightarrow c = z_{\alpha/2}$

Thus, DR: Reject H_0 if $|W| > z_{\alpha/2}$.

- With this DR, $P(\text{Type I error}) = P(\text{reject } H_0 \mid H_0 \text{ true})$

$$= P(|W| > z_{\alpha/2}) = P(W < z_{\alpha/2}) + P(W > z_{\alpha/2})$$
$$= \frac{\alpha}{2} + \frac{\alpha}{2} = \alpha.$$

Notes:
1) W is measuring how far away is $\hat{\theta}$ from θ_0 in terms of se's when H_0 is true
i.e. if H_0 is true $\hat{\theta} - \theta_0 \simeq 0$ (for large n) since $\hat{\theta} \to \theta = \theta_0$
else $|\hat{\theta} - \theta_0| > 0$

2) But precision or variability of $\hat{\theta}$ is important.

eg. $\hat{\theta} - \theta_0 = 0.1$

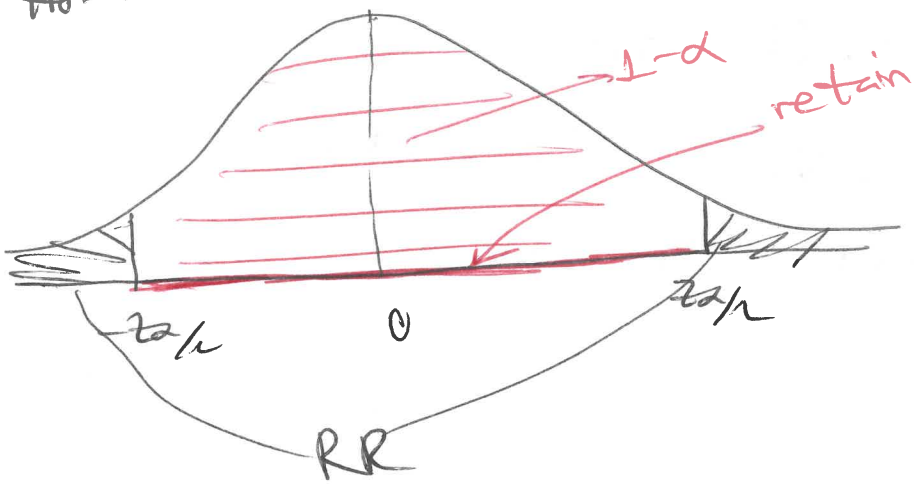
case (i) $se(\hat{\theta}) = 0.01 \longrightarrow W = \frac{0.1}{0.01} = 10$ strong evidence for H_a.

case (ii) $se(\hat{\theta}) = 100 \longrightarrow W = \frac{0.1}{100} = 0.001$ no evidence for H_a.
(highly consistent with H_0)

Thus, how many se's is the difference $\hat{\theta} - \theta_0$ is important.

- Connection (of Two-sided Tests) with $(1-\alpha)$ CI

If $H_0: \theta = \theta_0$ is true, then $W \sim Z \sim N(0,1)$



- Retain H_0 if $-z_{\alpha/2} < W < z_{\alpha/2}$

$$\Leftrightarrow -z_{\alpha/2} < \frac{\hat{\theta} - \theta_0}{se} < z_{\alpha/2}$$

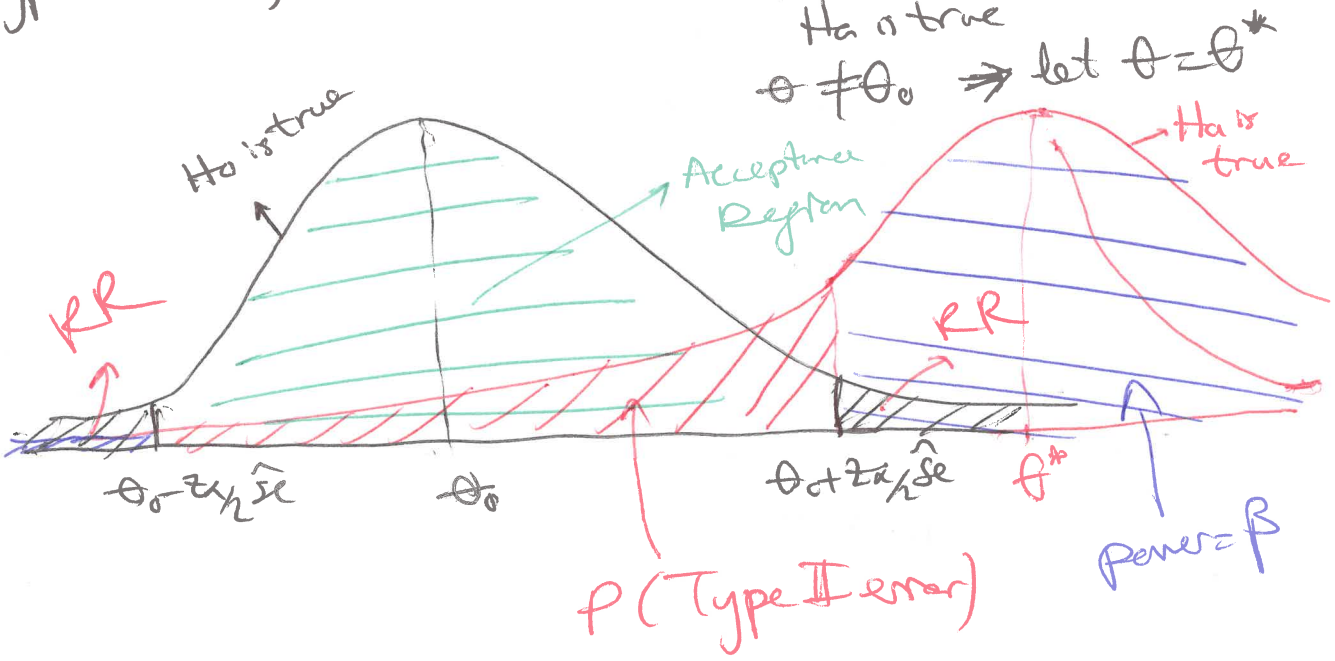
$$\Leftrightarrow \hat{\theta} - z_{\alpha/2} se < \theta_0 < \hat{\theta} + z_{\alpha/2} se$$

$$\Leftrightarrow \theta_0 \in \underbrace{(\hat{\theta} - z_{\alpha/2} se,\ \hat{\theta} + z_{\alpha/2} se)}_{(1-\alpha)\ CI}$$

So, repeat $H_0: \theta = \theta_0$, if θ_0 is not in $(1-\alpha)$ CI
retain H_0: " " " in the $(1-\alpha)$ CI.

$P(\text{Type II error}) = P(\text{Accept } H_0 \mid \underbrace{H_0 \text{ is false}}_{H_a \text{ is true}})$

$\theta \neq \theta_0 \Rightarrow$ let $\theta = \theta^*$



One-Sided Alternatives

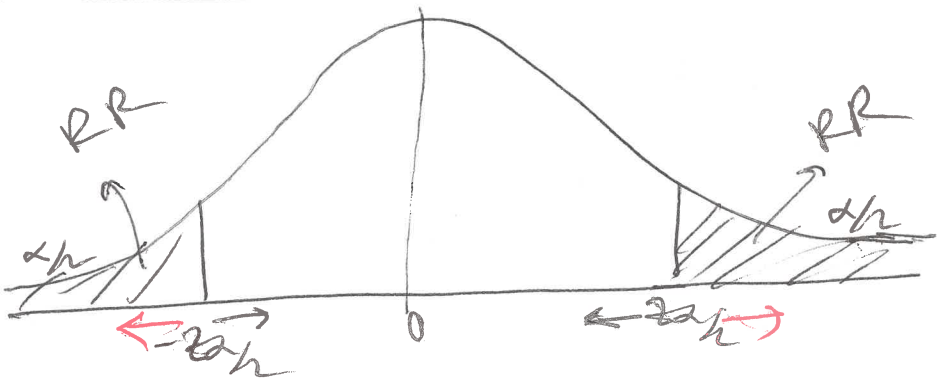
- If $H_0: \theta \leq \theta_0$ vs $H_a: \theta > \theta_0$

  DR: Reject H_0, if $W > z_\alpha$

- If $H_0: \theta \geq \theta_0$ vs $H_a: \theta < \theta_0$

  DR: Reject H_0, if $W < -z_\alpha$.

Dependence of 2-sided test on α



- If $\alpha \uparrow$, RR $\uparrow$
- If $\alpha \downarrow$, RR $\downarrow$

$P(\text{Type I error}) = \alpha$ "our tolerance prob. for rejecting a true H_0"

common choices $\alpha = 0.01$, (0.05) or 0.10

Ex: 64 tosses of a coin, 25 H, 39 T.

Test whether coin is fair or not at $\alpha = 0.05$ level.

Soln: $H_0: p = 0.5$ vs $H_a: p \neq 0.5$

$\hat{p} = \bar{X} = \frac{\sum X_i}{n} = \frac{25}{64}$ where $X_i = I(\text{coin shows H}) \sim \text{Bernoulli}(p)$ with $p = P(H)$,

$n = 64$ is large, so by CLT,

$\hat{p} \overset{\text{approx}}{\sim} \text{Normal}.\ (N(p, se^2(p)))$

So, Wald's test is applicable, with

test stat: $W = \frac{\hat{p} - p_0}{se(\hat{p})} = \frac{0.39 - 0.5}{se(\hat{p})}$ ↑ ?

$se(\hat{p}) = \sqrt{\frac{p(1-p)}{n}}$ but under $H_0: p = 0.5$

$se(\hat{p}) = \sqrt{\frac{0.5(1-0.5)}{64}} = \frac{1}{16} = 0.0625 \Rightarrow W = \frac{-0.11}{0.0625} = -1.76$

DR: Reject H_0 if $|W| > z_{0.025} = 1.96$

Here, we retain H_0, since $1.76 \not> 1.96$.

- one could also estimate $se(\hat{p})$ as $\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \cong 0.061$ (since $\hat{p}$ is consistent)

In fact if $H_0: \theta \leq 0.5$ or $H_0: \theta \geq 0.5$, we have to use this se.

Ex: What if we observe 10 H & 54 T in the previous example?

So, $W = \frac{\hat{p} - p_0}{se(\hat{p})} = \frac{\frac{10}{64} - 0.5}{\text{a.c.b.r.t}} = -5.5$

since $5.5 > 1.96$, we reject H_0.

Wald Test for Two Independent Populations

$D_1 = \{X_1, X_2 \longrightarrow X_n\}$ with $X_i \overset{iid}{\sim} F(x;\theta_1)$

$D_2 = \{Y_1, Y_2 \longrightarrow Y_m\}$ " $Y_j \overset{iid}{\sim} F(y;\theta_2)$

$X_i \perp Y_j$ for all i,j. & $\hat{\theta}_1$ & $\hat{\theta}_2$ are both asy. normal AN

HT: $H_0: \theta_1 = \theta_2$ vs $H_a: \theta_1 \neq \theta_2$

$\theta_1 = \theta_2 \Leftrightarrow \delta = \theta_1 - \theta_2 = 0$

$\Rightarrow H_0: \delta = 0 \overset{\delta_0}{\leftarrow}$ vs $H_a: \delta \neq 0$

and $\hat{\delta} = \hat{\theta}_1 - \hat{\theta}_2$ A.N. $\left(\hat{\theta}_1 - \hat{\theta}_2 \overset{approx}{\sim} N(\theta_1 - \theta_2, se^2(\hat{\theta}_1 - \hat{\theta}_2))\right)$

$$W = \frac{\hat{\delta} - \delta_0}{se(\hat{\delta})} = \frac{(\hat{\theta}_1 - \hat{\theta}_2) - 0}{se(\hat{\theta}_1 - \hat{\theta}_2)} = \frac{\hat{\theta}_1 - \hat{\theta}_2}{se(\hat{\theta}_1 - \hat{\theta}_2)} \Rightarrow |W| \overset{?}{>} z_{\alpha/2}$$

Here, $Var(\hat{\theta}_1 - \hat{\theta}_2) = Var(\hat{\theta}_1) + Var(\hat{\theta}_2)$

$= se^2(\hat{\theta}_1) + se^2(\hat{\theta}_2)$

$\Rightarrow se(\hat{\theta}_1 - \hat{\theta}_2) = \sqrt{se^2(\hat{\theta}_1) + se^2(\hat{\theta}_2)}$

Z-Test

Case (I)

$D = \{X_1, X_2, \ldots, X_n\}$, $X_i \overset{iid}{\sim} N(\mu, \sigma^2)$ μ unknown, σ^2 known.

$H_0: \mu = \mu_0$ vs $H_a: \mu \neq \mu_0$.

Under H_0, $Z = \dfrac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \sim N(0,1)$

DR: Reject H_0, if $|Z| > z_{\alpha/2}$.

Case II: D as above, $X_i \overset{iid}{\sim} F(x)$ where $E[X_i] = \mu$ unknown
$Var(X_i) = \sigma^2$ known

Then for large n ($n \geq 30$),

by CLT, $\bar{X} \overset{approx}{\sim} N(\mu, \frac{\sigma^2}{n})$

so, under H_0,

$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \overset{approx}{\sim} N(0,1)$$

DR: Reject H_0, if $|Z| > z_{\alpha/2}$

Note:

Wald test	vs	Z-test	
$W = \dfrac{\hat{\mu} - \mu_0}{se(\hat{\mu})}$		$Z = \dfrac{\hat{\mu} - \mu_0}{\sigma/\sqrt{n}}$	σ is assumed known
or $W = \dfrac{\hat{\mu} - \mu_0}{\hat{se}(\hat{\mu})}$		σ is known F is not known or non-normal, but $n \geq 30$ or $F \equiv$ Normal, any n.	

$\hat{\mu}$ is A.N.

- One-sided Z-tests

- $H_0: \mu \leq \mu_0$ vs $H_a: \mu > \mu_0$
  DR: Reject H_0 if $Z > z_\alpha$

- $H_0: \mu \geq \mu_0$ vs $H_a: \mu < \mu_0$
  DR: Reject H_0 if $Z < -z_\alpha$

Z-Test for 2 Indep. Populations

Case I: $D_1 = \{X_1, X_2, \ldots, X_n\}$, $X_i \overset{iid}{\sim} N(\mu_1, \sigma_1^2)$

$D_2 = \{Y_1, Y_2, \ldots, Y_m\}$, $Y_j \overset{iid}{\sim} N(\mu_2, \sigma_2^2)$ $\left(X_i \perp Y_j \text{ for all } i,j\right)$

μ_1, μ_2 unknown, σ_1^2 & σ_2^2 known

$H_0: \mu_1 = \mu_2$ vs $H_a: \mu_1 \neq \mu_2$

Let $\delta = \mu_1 - \mu_2 \Rightarrow H_0: \delta = \underbrace{0}_{\delta_0}$ vs $H_a: \delta \neq 0$

$$Z = \frac{(\bar{X} - \bar{Y}) - \delta_0}{se(\bar{X} - \bar{Y})} \leftarrow ?$$

Here, $Var(\bar{X} - \bar{Y}) = Var(\bar{X}) + Var(\bar{Y}) = \frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}$

$$\Rightarrow se(\bar{X} - \bar{Y}) = \sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}$$

DR: Reject H_0, if $|Z| > z_{\alpha/2}$

Case II $X_i \overset{iid}{\sim} F_X$ $Y_j \overset{iid}{\sim} F_Y$

w/ $E[X_i] = \mu_1$ $E[Y_j] = \mu_2$

$Var(X_i) = \sigma_1^2$ $Var(Y_j) = \sigma_2^2$ σ_1^2, σ_2^2 known

$n \geq 30$ $m \geq 30$

By CLT, $Z = \dfrac{(\bar{X} - \bar{Y}) - \delta_0}{\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}} \overset{\text{approx}}{\sim} N(0,1)$

DR: Reject H_0, if $|Z| > z_{\alpha/2}$.

t-Test

$D = \{X_1, X_2, \ldots, X_n\}$, $X_i \overset{iid}{\sim} N(\mu, \sigma^2)$ μ unknown, σ^2 "

$H_0: \mu = \mu_0$ vs $H_a: \mu \neq \mu_0$

$T = \dfrac{\bar{X} - \mu_0}{S/\sqrt{n}}$ where $S^2 = \dfrac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$

Under H_0, $T \sim t_{n-1}$

- **Notes**

1) If n is large ($n \geq 30$), by CLT, & Slutsky $\dfrac{\bar{X} - \mu_0}{S/\sqrt{n}} \overset{\text{approx}}{\sim} N(0,1)$

so, one can use a Z-test.

2) So, t-test is most useful when n is small ($n < 30$)

3) normality of X_i is a weak assumption, if not severely nonnormal t-test can still be used.

D.R. for t-test

Reject H_0, if $|T| > t_{\alpha/2, n-1}$ ← table look-up or use R, Python, etc.

One-Tailed t-Test

- $H_0: \mu \leq \mu_0$ vs $H_a: \mu > \mu_0$
  DR: Reject H_0, if $T > t_{\alpha, n-1}$
- $H_0: \mu \geq \mu_0$ vs $H_a: \mu < \mu_0$
  DR: Reject H_0, if $T < -t_{\alpha, n-1}$

t-Test for Two Indep. Populations (Unpaired)

$D_1 = \{X_1, X_2, \dots, X_n\}$, $X_i \overset{ind}{\sim} N(\mu_1, \sigma_1^2)$ · σ_1^2, σ_2^2 unknown

$D_2 = \{Y_1, Y_2, \dots, Y_m\}$, $Y_j \overset{ind}{\sim} N(\mu_2, \sigma_2^2)$ · $\min(n, m) < 30$

$X_i \perp Y_j$ for all i, j

$H_0: \mu_1 = \mu_2$ vs $H_a: \mu_1 \neq \mu_2$

$\delta = \mu_1 - \mu_2$

$$T = \frac{\bar{X} - \bar{Y} - \delta_0}{\widehat{se}(\bar{X} - \bar{Y})}$$

Case I: $\sigma_1^2 = \sigma_2^2 = \sigma^2$

$$S_p^2 = \frac{(n-1)S_1^2 + (m-1)S_2^2}{n+m-2} \quad \& \quad \widehat{se}(\bar{X} - \bar{Y}) = S_p \sqrt{\frac{1}{n} + \frac{1}{m}}$$

DR: Reject H_0, if $|T| > t_{\alpha/2, n+m-2}$

Case II: $\sigma_1^2 \neq \sigma_2^2$

$$\widehat{se}(\bar{X}-\bar{Y}) = \sqrt{\frac{S_1^2}{n} + \frac{S_2^2}{m}}$$

where $S_1^2 = \frac{1}{n-1}\sum (X_i - \bar{X})^2$

$S_2^2 = \frac{1}{m-1}\sum (Y_i - \bar{Y})^2$

DR: Reject H_0 if $|T| > t_{\alpha/2, df}$

where df is adjusted (Satterthwaite approximation)

$\max(n-1, m-1) \leq df \leq n+m-2$. (from software)

- One-tailed versions are possible with $t_{\alpha, df}$ or $-t_{\alpha, df}$ in DR.

If $H_0: \mu_1 \leq \mu_2$ vs $H_a: \mu_1 > \mu_2$

DR: Reject H_0 if $T > t_{\alpha, df}$.

Paired t-Test

$D_1 = \{X_1, X_2, \ldots, X_n\}$, $X_i \overset{iid}{\sim} N(\mu_1, \sigma_1^2)$ σ_1^2, σ_2^2 unknown.

$D_2 = \{Y_1, Y_2, \ldots, Y_n\}$, $Y_j \overset{iid}{\sim} N(\mu_2, \sigma_2^2)$

X_i & Y_i are dependent (there is reason to pair X_i & Y_i)

$H_0: \mu_1 = \mu_2$ vs $H_a: \mu_1 \neq \mu_2$

take differences

$D = \{d_1, d_2, \ldots, d_n\}$ where $d_i = X_i - Y_i$ $i = 1, 2, \ldots, n$

$d_i \overset{iid}{\sim} N(\mu_1 - \mu_2, \sigma_1^2 + \sigma_2^2) \equiv N(\delta, \sigma_\delta^2)$

$H_0: \delta = 0$ vs $H_a: \delta \neq 0$

$$T = \frac{\bar{d} - 0}{s_d/\sqrt{n}} = \frac{\bar{d}}{s_d/\sqrt{n}} \quad \text{where } \bar{d} = \frac{\sum d_i}{n} \text{ & } s_d^2 = \frac{1}{n-1}\sum (d_i - \bar{d})^2$$

DR: Reject H_0, if $|T| > t_{\alpha/2, n-1}$

p-Value (Observed Significance Level)

- An alternate way of reporting the result/decision of the HT.
- So far the conclusion was binary with only "reject H_0" or "retain H_0" at a given α level.
- Let T_{obs} be the observed (computed) value of the test stat.
- p-value: Probability of getting the current value of the statistic or more "extreme" values than the observed statistic

$$\text{p-value} = \text{Prob}(\text{statistic } T \text{ more extreme than } T_{obs} \mid H_0 \text{ is true})$$

"more extreme" means "towards" or "in favor of H_a" i.e. further "away from H_0."

DR (with p-value):

Reject H_0 if p-value $< \alpha$.

Ex: Recall Wald Test:

$D = \{X_1, X_2, \cdots, X_n\}$ $X_i \overset{iid}{\sim} F(x;\theta)$, $\hat{\theta}$ an estimator for θ

and $\hat{\theta}$ is A.N.

$$W = \frac{\hat{\theta} - \theta_0}{se(\hat{\theta})} \quad \text{or} \quad \frac{\hat{\theta} - \theta_0}{\hat{se}(\hat{\theta})}, \quad W \overset{approx}{\sim} N(0,1)$$

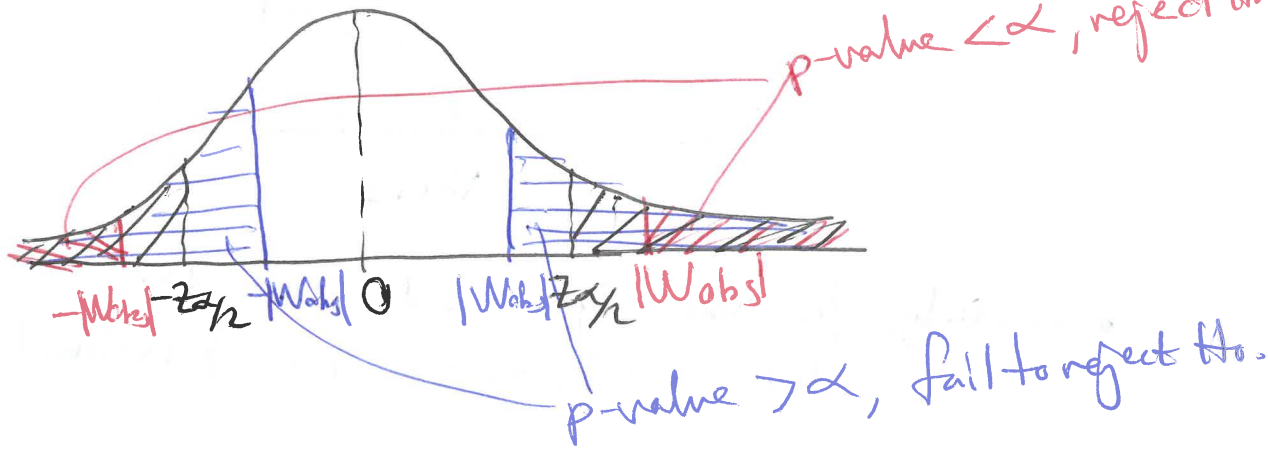
Let the observed or computed test stat be W_{obs}

Q: Which values of W would be more extreme than W_{obs} (under H_0)?

Ans. $|W| > |W_{obs}|$

so $p\text{-value} = P(|W| > |W_{obs}|)$

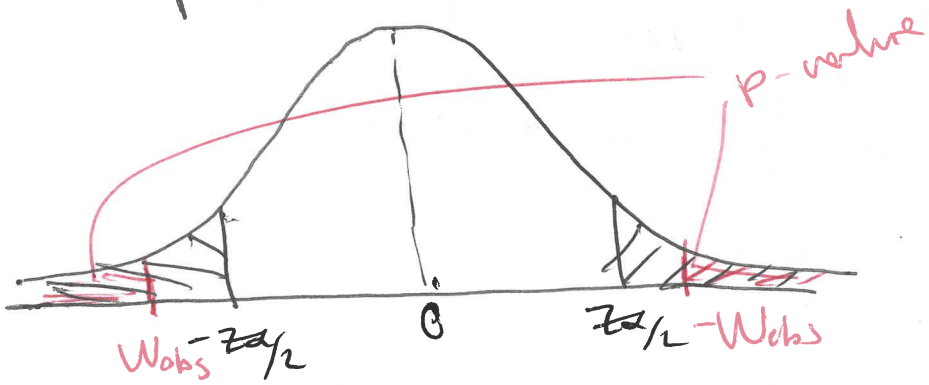
$= P(W > |W_{obs}|) + P(W < -|W_{obs}|)$



Since $W \overset{approx}{\sim} N(0,1)$,

$p\text{-value} \approx 2(1 - \Phi(|W_{obs}|)) = 2\Phi(-|W_{obs}|)$

Ex: If $W_{obs} < 0$, $p\text{-value} = 2\Phi(W_{obs})$



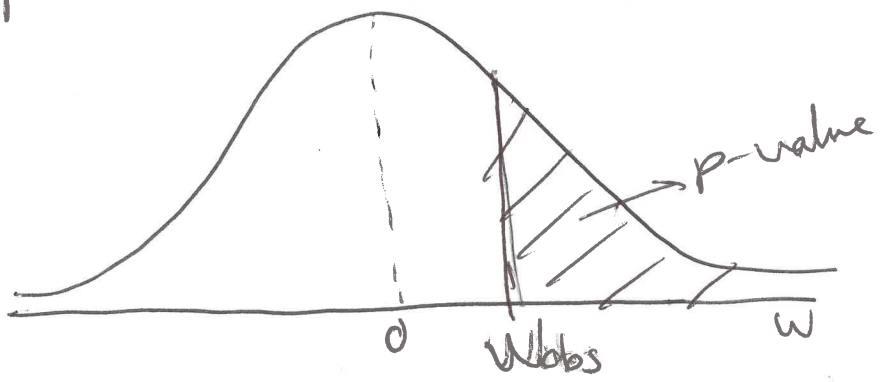
p-value for One-Sided Wald Test

$H_0: \theta \le \theta_0$ vs $H_a: \theta > \theta_0$, p-value = ? (right-sided)

$W = \frac{\hat{\theta} - \theta_0}{se(\hat{\theta})}$, $W_{obs} = \frac{\hat{\theta} - \theta_0}{se(\hat{\theta})}$

$W > W_{obs}$ would be more extreme W values than observed.

So, p-value $= P(W > W_{obs}) = 1 - \Phi(W_{obs})$



- Similarly, for the left-sided alternative,
  p-value $= P(W < W_{obs}) = \Phi(W_{obs})$

- p-values for Z-tests & t-tests are similar.

Ex: For 2-sided Z-test with $Z_{obs} > 0$

$$\text{p-value} = P(Z > Z_{obs}) + P(Z < -Z_{obs})$$
$$= 2(1 - \Phi(Z_{obs}))$$

- p-values for one-sided Z-tests can also be obtained by replacing W, W_{obs} with Z, Z_{obs}, respectively.

- Likewise for the t-tests.

Two Caveats: (1) A large p-value is not strong evidence in favor of H_0. " " can occur when (i) H_0 is true or (ii) H_0 is false but test has low power

(2) p-value $\neq$ Prob(H_0 is true | data)

(Comparing correct prediction rates of 2 prediction algorithms)

Ex: $D_1 = \{X_1, X_2, \ldots, X_n\}$, $X_i \overset{iid}{\sim} \text{Bernoulli}(p_1)$

$D_2 = \{Y_1, Y_2, \ldots, Y_m\}$, $Y_j \overset{iid}{\sim} \text{Bernoulli}(p_2)$ $\left(X_i \perp Y_j \text{ for all } i,j\right)$

Is $p_1 = p_2$? (n, m are large)

Soln: $H_0: p_1 = p_2$ vs $H_a: p_1 \neq p_2$

equivalently

$H_0: p_1 - p_2 = 0$ vs $H_a: p_1 - p_2 \neq 0$

The MLE of $p_1 - p_2$ is $\hat{p}_1 - \hat{p}_2$, since

" " of p_1 is $\hat{p}_1$ &

" " " p_2 " $\hat{p}_2$.

$\left.\begin{array}{l} \hat{p}_1 \overset{\text{approx}}{\sim} N(p_1, \text{Var}(\hat{p}_1)) \\ \hat{p}_2 \overset{\text{approx}}{\sim} N(p_2, \text{Var}(\hat{p}_2)) \end{array}\right\} \Rightarrow \hat{p}_1 - \hat{p}_2 \overset{\text{approx}}{\sim} N(p_1 - p_2, \text{Var}(\hat{p}_1) + \text{Var}(\hat{p}_2))$

where $\text{Var}(\hat{p}_1) = \frac{p_1(1-p_1)}{n}$ & $\text{Var}(\hat{p}_2) = \frac{p_2(1-p_2)}{m}$

So, Wald test stat is

$$W = \frac{(\hat{p}_1 - \hat{p}_2) - 0}{se(\hat{p}_1 - \hat{p}_2)} \quad \text{where} \quad se(\hat{p}_1 - \hat{p}_2) = \sqrt{\text{Var}(\hat{p}_1) + \text{Var}(\hat{p}_2)}$$

$$= \sqrt{\frac{p_1(1-p_1)}{n} + \frac{p_2(1-p_2)}{m}}$$

But p_1, p_2 unknown

$\hat{p}_1$ is constant for p_1 &

$\hat{p}_2$ " " " p_2 $\Rightarrow \hat{se}(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n} + \frac{\hat{p}_2(1-\hat{p}_2)}{m}}$

$$\Rightarrow W = \frac{\hat{p}_1 - \hat{p}_2}{\hat{se}(\hat{p}_1 - \hat{p}_2)}$$

D.R. Reject H_0, if $|W| > z_{\alpha/2}$.

OR compute p-value $= 2 \min\left(P(W > W_{obs}), P(W < W_{obs})\right)$

DR: Reject H_0, if p-value $< \alpha$.

Likelihood Ratio Test (LRT)

$H_0: \theta \in \Omega_0$ vs $H_a: \theta \in \Omega \setminus \Omega_0$

LRT stat is

$$T_{LRT} = 2 \log\left(\frac{\sup_{\theta \in \Omega} L(\theta)}{\sup_{\theta \in \Omega_0} L(\theta)}\right)$$ where $L(\theta) = f(x_1, x_2, \ldots, x_n; \theta)$ (likelihood; joint pdf)

$$= 2 \log\left(\frac{L(\hat{\theta}_{MLE})}{L(\hat{\theta}_0)}\right)$$

where $\hat{\theta}_{MLE}$ is the (global) MLE &
$\hat{\theta}_0$ is " MLE under H_0.

DR: Reject H_0 if $T_{LRT} > c$ ← critical value based on distr. of T_{LRT}.

Note: Almost all tests we have seen are special cases of LRT (i.e. can be derived from T_{LRT}).