# §8.1: Hypotheses and Test Procedures

**Motivating Example:** You are given a coin to flip. Flipping the coin 100 times, you get "Heads" 60 times. **Question:** Is the coin fair?

The probability of getting heads 60 or more times (if the coin were fair) is

$P(X \geq 60)$ where $X \sim \text{Binomial}(100, \frac{1}{2})$

$= 1 - \text{pbinom}(59, 100, \frac{1}{2}) \approx .028$

This probability is "fairly low" so you conclude that the coin is likely unfair.

**Note:** You could be more certain if you flip the coin more times.

- 120 H out of 200 flips $\Rightarrow$ Prob $\approx .004$
- 180 H out of 300 flips $\Rightarrow$ Prob $\approx .0007$

**Alternate view.** If the coin were fair then we expect $E[X] = 50$ Heads.

The 90% Confidence Interval around 50 is

$(\text{qbinom}(.05, 100, \frac{1}{2}), \text{qbinom}(.95, 100, \frac{1}{2}))$

$= (42, 58)$

---

**Def:** A statistical hypothesis "H" is a statement about the distribution of a random variable

[usually about the value of a parameter

Ex: "$H: \theta = \theta_0$" ← some number]

A test statistic is the function of sample data used to "test" the hypothesis

[usually this is an estimator for the parameter: $\hat{\theta}$ gives $\hat{\theta} \approx \theta$]

---

**Example:** To test hypothesis "$H: \mu = 10$"

We would use sample mean as test statistic

$\bar{x} \approx \mu$

---

**Hypothesis Testing** is a standard procedure for deciding the validity (or likelihood) of a statistical hypothesis, using observed data (the test statistic)

This always involves two hypotheses:

- $H_0$ "The Null Hypothesis"
- $H_A$ "The Alternative Hypothesis"

$H_0$: the "null hypothesis" is the default or "status quo" hypothesis. It describes the simplest possible situation — "similar populations are identical" or "everything is independent."

[Following the principle of "Occam's Razor" if the Null Hypothesis is possible then it is assumed.]

$H_A$: the "alternative hypothesis" is everything which is not the null hypothesis — including any value $\hat{\theta}$ which statistical sampling suggests.

The goal of classical hypothesis testing is to show that sampled data is incompatible with null hyp.

"Reject the null hypothesis" — if the null hyp. is not true, then the default new assumption is $\hat{\theta}$ (whatever the sampling data suggests)

Results of hypothesis test are either

→ "Reject the null hypothesis"

→ "Fail to reject the null hypothesis"

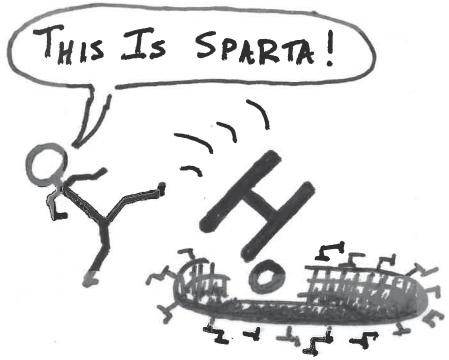
We will divide discussion of hypothesis testing into two parts:

- Test Procedure
- Test Design

Note: This is separate from issues of experiment procedure & design.

(For example: proper use of control group, blind trials, avoiding confounding variables, considering correlation vs causation, etc...)

In this set of notes, I will give a series of examples of hypothesis test procedure.




"Reject the Null Hypothesis"


"Fail to reject the Null Hypothesis"

Hypothesis Testing problems have two different types:

"Two-Tailed"

$$\begin{bmatrix} H_0: \theta = \theta_0 \\ H_A: \theta \neq \theta_0 \end{bmatrix}$$

"One-Tailed"

$$\begin{bmatrix} H_0: \theta \leq \theta_0 \\ H_A: \theta > \theta_0 \end{bmatrix} \text{ or } \begin{bmatrix} H_0: \theta \geq \theta_0 \\ H_A: \theta < \theta_0 \end{bmatrix}$$

(Depending on whether $\hat{\theta} > \theta_0$ or $\hat{\theta} < \theta_0$)

Look at estimate $\hat{\theta}$ from sample data

**Def:** The p-value of a test statistic is the probability of obtaining sample data at least as extreme as observed value assuming $H_0$.

Two-Tailed

$$\text{p-value} = P\left( |\hat{\Theta} - \theta_0| > |\hat{\theta} - \theta_0| \;\middle|\; H_0 \text{ is true} \right)$$

($H_0$ is true $\Leftrightarrow \theta = \theta_0$)

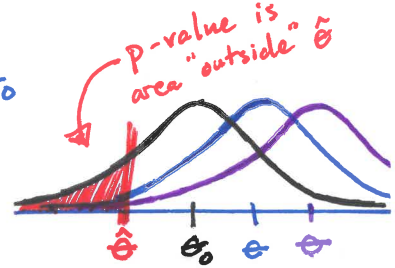
One-Tailed

$$\text{p-value} = P\left( \hat{\Theta} - \theta_0 > \hat{\theta} - \theta_0 \;\middle|\; H_0 \text{ is true} \right)$$

— or —

$$= P\left( \hat{\Theta} - \theta_0 < \hat{\theta} - \theta_0 \;\middle|\; H_0 \text{ is true} \right)$$

**Note:** In one-tailed case $\theta \geq \theta_0$ the maximum p-value occurs when $\theta = \theta_0$. So we use this assumption here, too.




## §8.2 "z-Tests": Testing $\mu = E[X]$

Suppose $X$ is Normal with $\sigma$ known

— or —

$X$ is maybe not Normal, $\sigma$ unknown but #samples is large ($n > 40$)

Recall: $\dfrac{\bar{X} - \mu}{s/\sqrt{n}} \approx \text{Normal}(0, 1)$

Sample $X$ to get sample mean $\bar{x}$.

Test against null hypothesis $\mu = \mu_0$

(Usually $\mu_0$ is the expected value for a similar population or is 0)

$H_0: \mu = 0 \iff$ "$X$ is random noise"

Test statistic is $\bar{x}$ ← Point estimate for $\mu = E[X]$

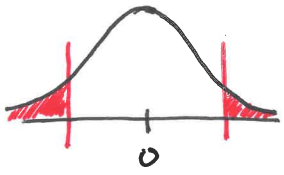
z-score is #(std dev) of $\bar{x}$ from $\mu_0$

$$z = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

Example: If z-score is 1.5 then $\bar{x}$ is 1.5 standard deviations away from $\mu_0$

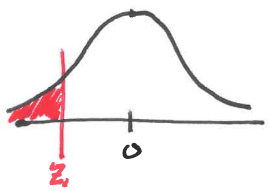
p-value is $P(|Z| > |z|)$ two-tailed

$\left\{ \begin{matrix} P(Z < z) \\ P(Z > z) \end{matrix} \right\}$ one-tailed

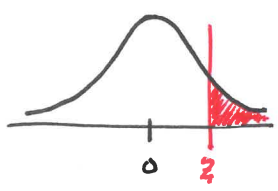



$P(|Z| > |z|)$

Two-Tailed Test




$P(Z < z_1)$




$P(Z > z_2)$

One-Tailed Tests

**Note:** Because Normal distribution is symmetric,

$$\left(\begin{array}{c}\text{p-value for}\\ \text{Two-Tailed Test}\end{array}\right) = 2\left(\begin{array}{c}\text{p-value for}\\ \text{One-Tailed Test}\end{array}\right)$$

Reject null hypothesis if p-value is less than some pre-determined cutoff $\alpha$ ("significance level")

$\alpha$ is usually 5%, 1%, or 0.5%

It is a standard value determined by

- Journal where you are publishing results
- Conventions set by other researchers in your field of study.

Smaller $\alpha$ are better but they usually will require larger (& more expensive) experiments...

**Example:** Sample $X$ 100 times and compute sample mean $\bar{x} = 21.4$ and sample std. dev $s = 6$

A similar population has $\mu_0 = 20$.

Hypothesis test against $\mu = 20$.

$$\text{z-score: } z_1 = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{21.4 - 20}{6/\sqrt{100}}$$

$$\approx 2.33$$

**Two Tailed Test**

$H_0: \mu = 20$

$H_A: \mu \neq 20$

p-value

$$p = P(|Z| > 2.33)$$

$$= 2 \cdot \text{pnorm}(2.33)$$

$$\approx 0.0198$$


**One Tailed Test**

$H_0: \mu \leq 20$

$H_A: \mu > 20$

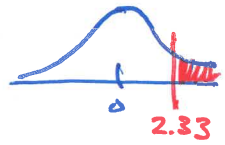
Note: $\bar{x} = 21.4 > 20$

p-value

$$p = P(Z > 2.33)$$

$$= [1 - \text{pnorm}(2.33)]$$

$$\approx 0.0099$$




Note: also equal to pnorm(-2.33)

Two tailed test will reject $H_0$ if $\alpha = .05$

fail to reject $H_0$ if $\alpha = .01$

One tailed test will reject $H_0$ even if $\alpha = .01$

## §8.3 "t-Tests" : Testing $\mu = E[X]$

Suppose $X$ is Normal, $\sigma$ unknown and #samples is not large

**Recall:** $\dfrac{\bar{X} - \mu}{S/\sqrt{n}} \sim t(n-1)$

(# samples $= n \geq 2 \Rightarrow$ degrees of freedom $= n-1 \geq 1$)

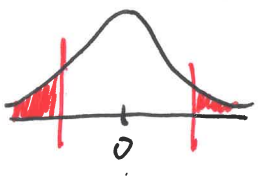
(In practice people will often use this if $X$ is Normal even if $n$ is big...)

(Test against $\mu = \mu_0$)

Test statistic is $\bar{x}$

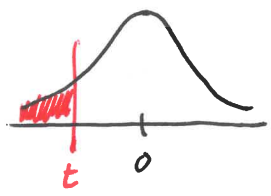
"t-score" is $\dfrac{\bar{x} - \mu_0}{S/\sqrt{n}}$ with $(n-1)$ deg. of freedom

p-value is




$P(|T| > |t|)$

Two-Tailed Test





$P(T < t)$ or $P(T > t)$

One-Tailed Tests

**Example** From Turkish government data the average height of women in Turkey in 2003 was 156.4 cm.

Suppose we measure height of 20 female students and get average height of 160 cm with std. dev. 6 cm.

Hypothesis test against $\mu = 156.4$ cm.

"Is the difference in avg height we found statistically significant?"

t-score $t = \dfrac{\bar{x} - \mu_0}{S/\sqrt{n}} = \dfrac{160 - 156.4}{6/\sqrt{20}}$

$\approx 2.68$ with 19 degrees of freedom

### Two-Tailed Test

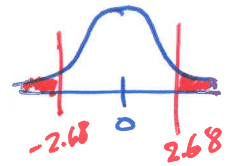
$H_0: \mu = 156.4$
$H_A: \mu \neq 156.4$

p-value

$p = P(|T| > 2.68)$

$= 2 \cdot pt(2.68, 19)$

$\approx 0.0148$




fail to reject if $\alpha = .01$

### One-Tailed Test

$H_0: \mu \leq 156.4$
$H_A: \mu > 156.4$

Note: $\bar{x} = 160 > 156.4$

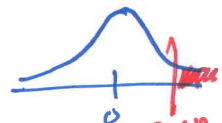
p-value

$p = P(T > 2.68)$

$= [1 - pt(2.68, 19)]$

$\approx 0.0074$

$= pt(-2.68, 19)$



## §8.4 Population Proportion (Binomial or "z")

Suppose we sample a large population and count number of occurences of something.

Example: Count # smokers
or # football fans

$X \sim \text{Binomial}(n, p)$

[ n = size of sample
p = proportion $\frac{\#\text{occurences}}{\#\text{population}}$ ]

Recall: If n is big & $p, q \geq 10/n$ ($q = 1-p$ "complementary proportion")
then
$\hat{p} = X/n \approx \text{Normal}\left(p, \sqrt{pq/n}\right)$

Two ways to test against "$H_0: p = p_0$".
→ Binomial
→ Normal

(Note: If the population we are sampling from is not that big, then $X \sim$ HyperGeometric...)

Example: 40% of men in Turkey smoke. Suppose we sample 100 male students at METU-NCC and find 30 smokers. Test against
$H_0$: Smokers at METU-NCC is same proportion as Turkey. $p = .4$

Expected # smokers in METU-NCC sample is $100(.4) = 40$ ← found $30 < 40$.
We will do one-tailed tests $H_A: p < .4$

**Binomial** $H_0: X \sim \text{Binomial}(100, .4)$

p-value $p = P(X \leq 30)$
$= \text{pbinom}(30, 100, .4)$
$\approx .0248$ ← Reject $H_0$ if $\alpha = .05$
Fail to reject if $\alpha = .01$

**Normal** $H_0: X/100 \approx \text{Normal}\left(.4, \sqrt{\frac{(.4)(.6)}{100}}\right)$

z-score $z = \frac{X/100 - .4}{\sqrt{\frac{(.4)(.6)}{100}}} = \frac{.3 - .4}{\sqrt{\frac{(.4)(.6)}{100}}}$

$\approx -2.04$

p-value $p = P(Z < -2.04)$
$= \text{pnorm}(-2.04) \approx .0206$

Not quite as accurate as Binomial, but not too far off.