Hypothesis Testing

Prof. Pradeep Ravikumar
pradeepr@cs.cmu.edu

Motivation

- Any data analysis algorithm applied to a set of data will produce some result(s)
  - There have been claims that the results reported in more than 50% of published papers are false (Ioannidis, 2005)
- Results may be a result of random variation
  - Any particular data set is a finite sample from a larger population
  - Often significant variation among instances in a data set
  - Unusual events or coincidences do happen, especially when looking at lots of events
  - For this and other reasons, results may not replicate, i.e., generalize to other samples of data
- Data scientists need to help ensure that results of data analysis are not **false discoveries**, i.e., not meaningful or reproducible

Significance & Hypothesis Testing

- Testing approaches are used to help avoid many of these problems

- Ultimate verification lies in the real world

Testing

- Make inferences (decisions) about that validity of a result
- For this, we need two things:
  - A statement that we want to disprove
    - Called the **null hypothesis (H_0)**
    - The null hypothesis is typically a statement that the result is merely due to random variation
    - It is the opposite of what we would like to show
  - A random variable, R, called a **test statistic**
    - The distribution of R under H_0 is called the **null distribution**
    - The value of R is obtained from the result and is typically numeric

Examples of Null Hypotheses

- A coin or a die is a fair coin or die.
- The difference between the means of two samples is 0
- The purchase of a particular item in a store is unrelated to the purchase of a second item, e.g., the purchase of bread and milk are unconnected

Significance Testing

- Significance testing was devised by the famous statistician Ronald Fisher

- For many years, significance testing has been a key approach for justifying the validity of scientific results

- Introduced the concept of **p-value**, which is widely used and misused

How Significance Testing Works

- Analyze the data to obtain a result
  - For example, data could be from flipping a coin 10 times to test its fairness
- The result is expressed as a value of the test statistic, R
  - For example, let R be the number of heads in 10 flips
- Compute the probability of seeing the current value of R or something more extreme
  - This probability is known as the **p-value** of the test statistic

How Significance Testing Works …

- If the p-value is sufficiently small, we say that the result is statistically significant
  - We say we reject the null hypothesis, H_0
  - A threshold on the p-value is called the **significance level, α**
    - Often the significance level is 0.01 or 0.05
- If the p-value is **not** sufficiently small, we say that we fail to reject the null hypothesis
  - Sometimes we say that we accept the null hypothesis, but a high p-value does not necessarily imply the null hypothesis is true

$$\text{p-value} = P(R|H_0) \neq P(H_0|R) = \frac{P(R|H_0)\,P(H_0)}{P(R)}$$

Example: Testing a coin for fairness

- H_0: $P(X=1) = P(X=0) = 0.5$
- Define the test statistic R to be the number of heads in 10 flips
- Set the significance level α to be 0.05
- The number of heads R has a binomial distribution
- For which values of R would you reject H_0?

k	$P(S=k)$
0	0.001
1	0.01
2	0.044
3	0.117
4	0.205
5	0.246
6	0.205
7	0.117
8	0.044
9	0.01
10	0.001

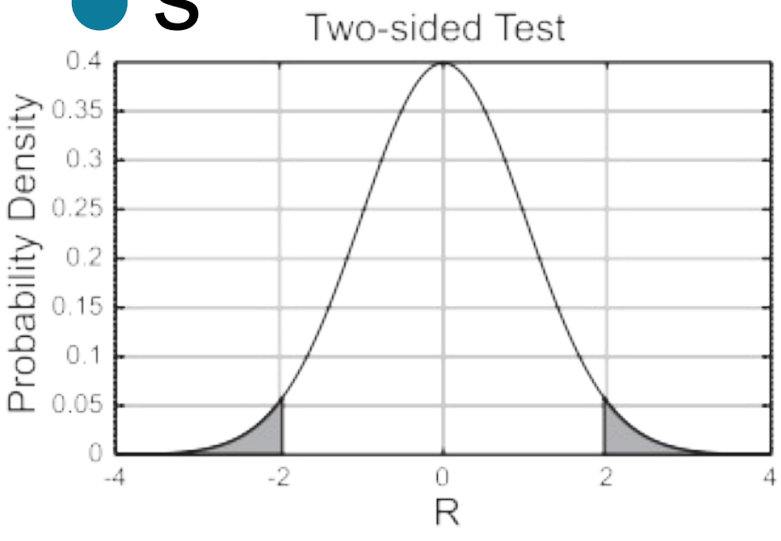
One-sided and Two-sided Tests

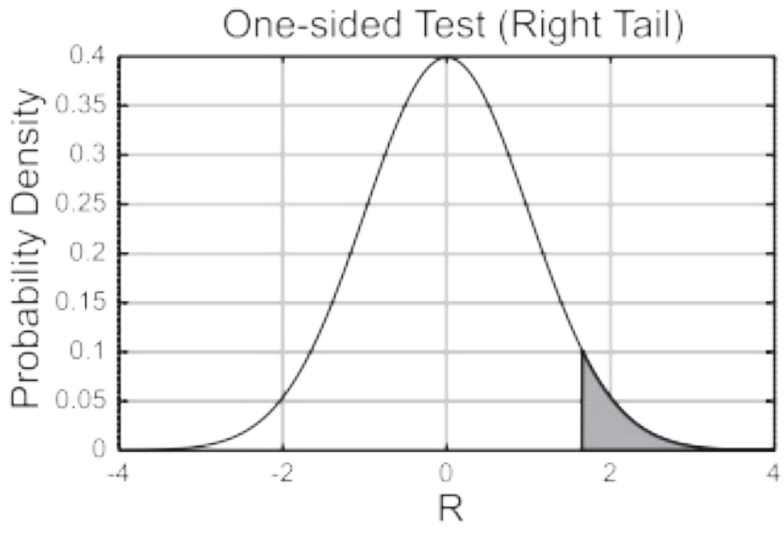
- More extreme can be interpreted in different ways
- For example, an observed value of the test statistic, R_{obs}, can be considered extreme if
  - it is greater than or equal to a certain value, R_H ,
  - smaller than or equal to a certain value, R_L, or
  - outside a specified interval, $[R_H, R_L]$.
- The first two cases are “one-sided tests” (right-tailed and left-tailed, respectively),
- The last case results in a “two-sided test.”

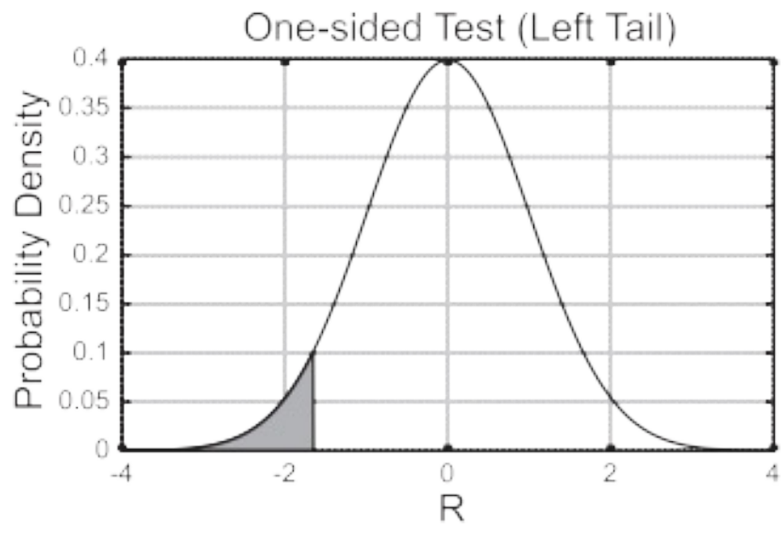
One-sided and Two-sided Tests …

- Example of one-tailed and two tailed tests for a test statistic R that is normally distributed for a roughly 5% significance level.

- S

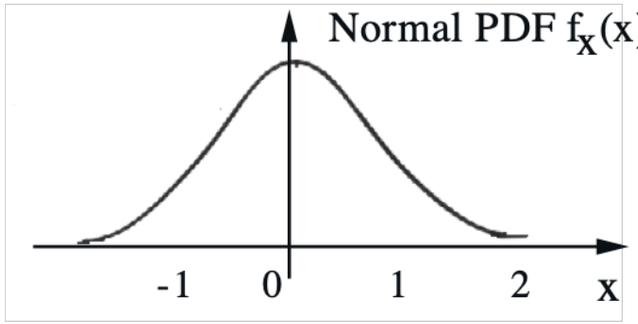






Normal Distribution

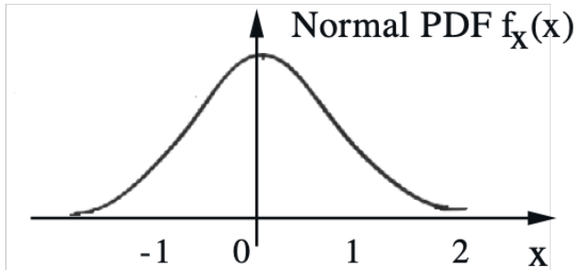
- Standard normal $N(0,1)$: $f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$



- $\mathbf{E}[X] = 0$ $\quad$ $\mathrm{var}(X) = 1$

Normal Distribution

- Standard normal $N(0,1)$: $f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$



- $\mathbf{E}[X] = 0 \qquad \text{var}(X) = 1$
- General normal $N(\mu, \sigma^2)$:

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$$

- It turns out that:
  $\mathbf{E}[X] = \mu$ and $\text{Var}(X) = \sigma^2$.

Background Recap: Normal Distribution

- Let $Y = aX + b$ where $X \sim N(\mu, \sigma^2)$

  Then: $\mathbf{E}[Y] = a\mathbf{E}[X] + b = a\mu + b, \quad \mathrm{Var}(Y) = a^2\mathrm{Var}(X) = a^2\sigma^2$

Normal Prob.

- Let $Y = aX + b$ where $X \sim N(\mu, \sigma^2)$

  Then: $\mathbf{E}[Y] = a\mathbf{E}[X] + b = a\mu + b, \quad \mathrm{Var}(Y) = a^2\mathrm{Var}(X) = a^2\sigma^2$

  - Fact: $Y \sim N(a\mu + b, a^2\sigma^2)$

Normal Prob.

- Let $Y = aX + b$ where $X \sim N(\mu, \sigma^2)$

  Then: $\mathbf{E}[Y] = a\mathbf{E}[X] + b = a\mu + b, \quad \mathrm{Var}(Y) = a^2\mathrm{Var}(X) = a^2\sigma^2$

  - Fact: $Y \sim N(a\mu + b, a^2\sigma^2)$

- If $X \sim N(\mu, \sigma^2)$, then $\dfrac{X - \mu}{\sigma} \sim N($ $)$

Normal Prob.

- Let $Y = aX + b$ where $X \sim N(\mu, \sigma^2)$

  Then: $\mathbf{E}[Y] = a\mathbf{E}[X] + b = a\mu + b, \quad \mathrm{Var}(Y) = a^2\mathrm{Var}(X) = a^2\sigma^2$

  - Fact: $Y \sim N(a\mu + b, a^2\sigma^2)$

- If $X \sim N(\mu, \sigma^2)$, then $\dfrac{X - \mu}{\sigma} \sim N(\quad 0 \ , \ 1 \quad)$

Calculating Normal Probabilities

- No closed form available for CDF
  - but there are tables (for standard normal)
  - $\mathbf{P}(Z \leq z)$, for different values of z.

	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319

Calculating Normal Probabilities

- No closed form available for CDF
  - but there are tables (for standard normal)
  - $\mathbf{P}(Z \leq z)$, for different values of z.

	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319

- If $X \sim N(\mu, \sigma^2)$, then $\dfrac{X-\mu}{\sigma} \sim N(\ 0\ ,\ 1\ )$
- If $X \sim N(2, 16)$:

$$\mathbf{P}(X \leq 3) = \mathbf{P}\left(\frac{X-2}{4} \leq \frac{3-2}{4}\right) = \text{CDF}(0.25)$$

Calculating Normal Probabilities

- No closed form available for CDF
  - but there are tables (for standard normal)
  - $\mathbf{P}(Z \leq z)$, for different values of z.

	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319

- Suppose we want the value of z_α s.t. $\mathbf{P}(|Z| \leq z_\alpha) = 1 - \alpha$.
- Equivalent to: $\mathbf{P}(Z \leq z_\alpha) = 1 - \alpha/2$
- Lookup value of z_α from table corresponding to prob. of $1 - \alpha/2$

Neyman-Pearson Hypothesis Testing

- Devised by statisticians Neyman and Pearson in response to perceived shortcomings in significance testing
  - Explicitly specifies an **alternative hypothesis, H_1**
  - Significance testing cannot quantify how an observed results supports H_1
  - Define an **alternative distribution** which is the distribution of the test statistic if H_1 is true
  - We define a **critical region** for the test statistic R
    - If the value of R falls in the critical region, we reject H_0
    - We may or may not accept H_1 if H_0 is rejected
  - The **significance level, α,** is the probability of the critical region under H_0

Hypothesis Testing …

- **Type I Error (α)**: Error of incorrectly rejecting the null hypothesis for a result.
  - It is equal to the probability of the critical region under H_0, i.e., is the same as the significance level, $\boldsymbol{\alpha}$.
  - Formally, $\alpha = P(R \in \text{Critical Region} \mid H_0)$

- **Type II Error (β)**: Error of falsely calling a result as not significant when the alternative hypothesis is true.
  - It is equal to the probability of observing test statistic values outside the critical region under H_1
  - Formally, $\beta = P(R \notin \text{Critical Region} \mid H_1)$.

Hypothesis Testing …

- Power: which is the probability of the critical region under H_1, i.e., $1-\beta$.
  - Power indicates how effective a test will be at correctly rejecting the null hypothesis.
  - Low power means that many results that actually show the desired pattern or phenomenon will not be considered significant and thus will be missed.
  - Thus, if the power of a test is low, then it may not be appropriate to ignore results that fall outside the critical region.

Example: Classifying Medical Results

- The value of a blood test is used as the test statistic, R, to identify whether a patient has a particular disease or not.
  - H_0: For patients **not** having the disease, R has distribution $\mathcal{N}(40, 5)$
  - H_1: For patients having the disease, R has distribution $\mathcal{N}(60, 5)$

Example: Classifying Medical Results

- The value of a blood test is used as the test statistic, R, to identify whether a patient has a particular disease or not.
  - H_0: For patients **not** having the disease, R has distribution $\mathcal{N}(40, 5)$
  - H_1: For patients having the disease, R has distribution $\mathcal{N}(60, 5)$

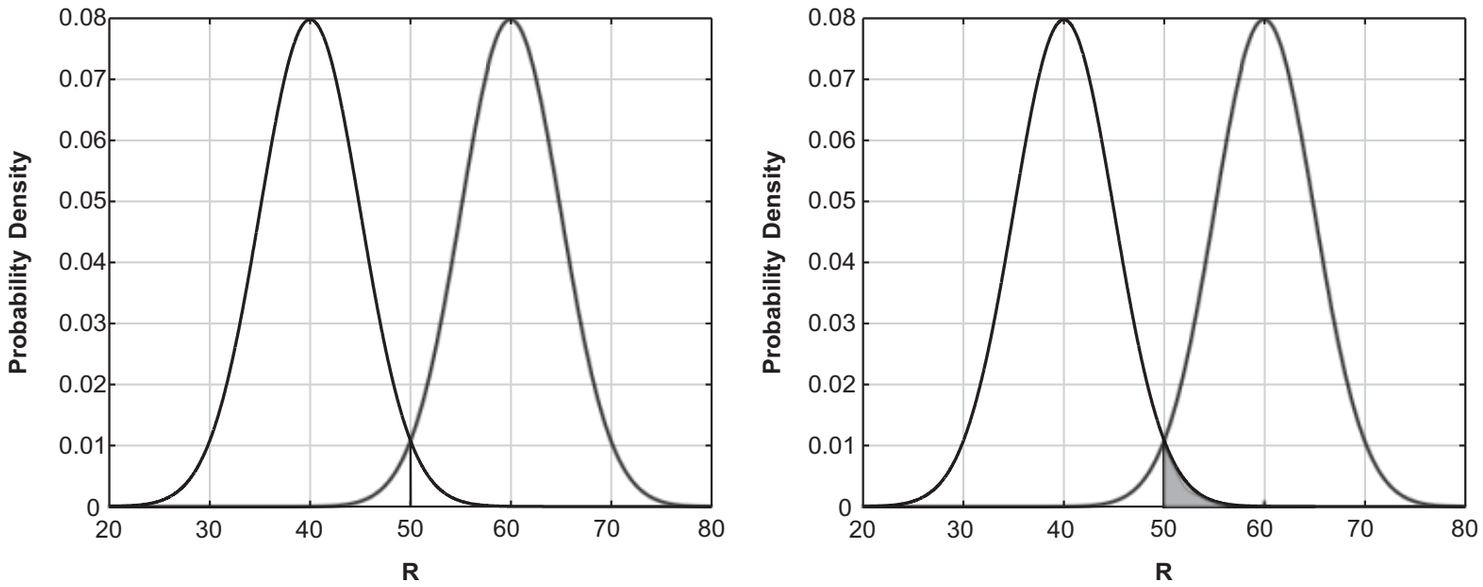
$$\alpha = \int_{50}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(R-u)^2}{2\sigma^2}} \, dR = \int_{50}^{\infty} \frac{1}{\sqrt{50\pi}} e^{-\frac{(R-40)^2}{50}} \, dR = 0.023, \mu = 40, \sigma = 5$$

$$\beta = \int_{-\infty}^{50} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(R-u)^2}{2\sigma^2}} \, dR = \int_{-\infty}^{50} \frac{1}{\sqrt{50\pi}} e^{-\frac{(R-60)^2}{50}} \, dR = 0.023, \mu = 60, \sigma = 5$$

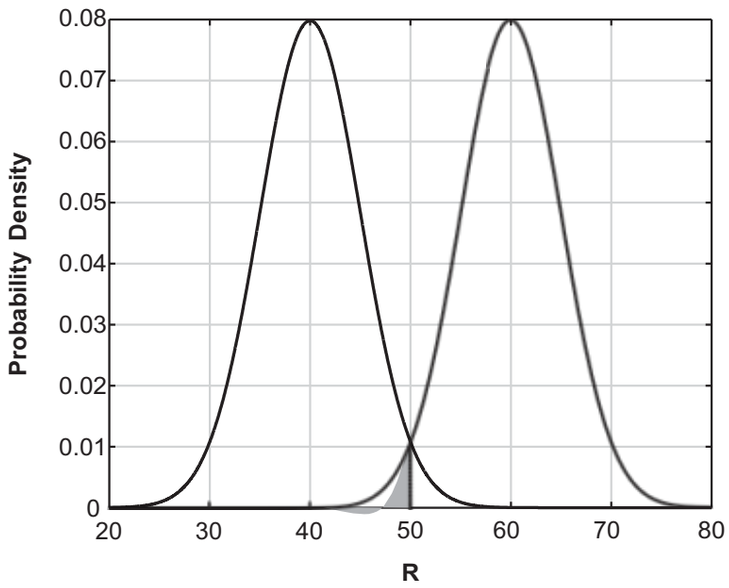
Power = 1 - β = 0.977

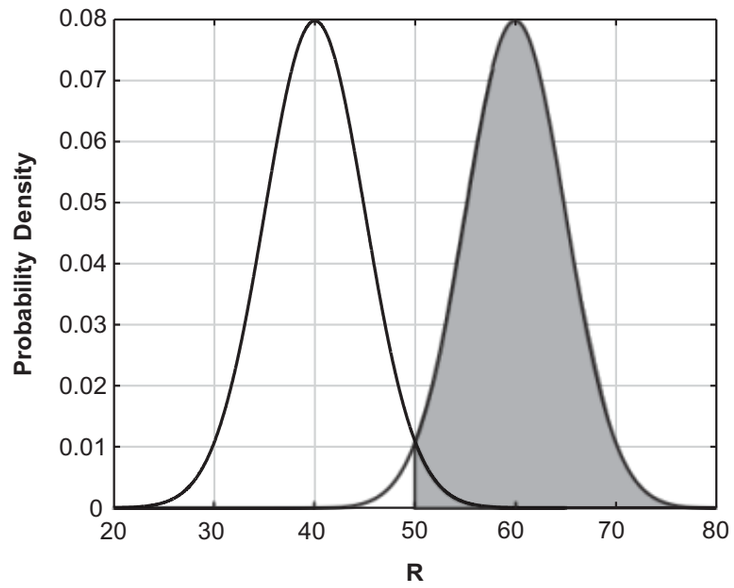
- See figures on the next page

α, β and Power for Medical Testing Example



Distribution of test statistic for the alternative hypothesis (rightmost density curve) and null hypothesis (leftmost density curve). Shaded region in right subfigure is α.

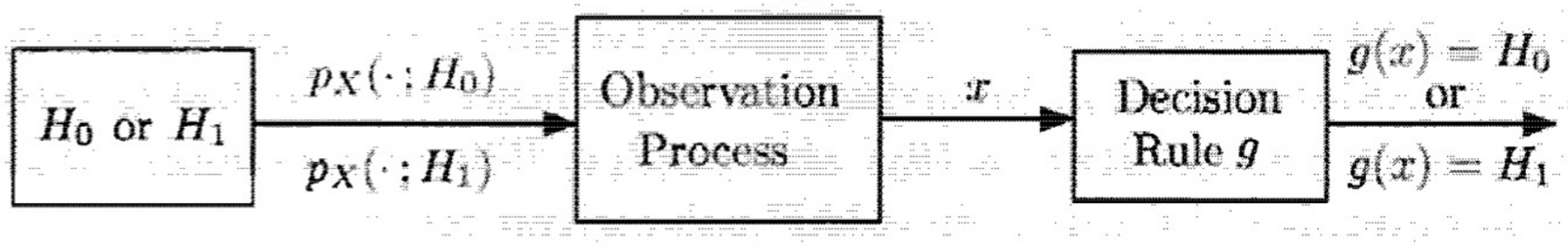




Shaded region in left subfigure is β and shaded region in right subfigure is power.

Binary Hypothesis Testing

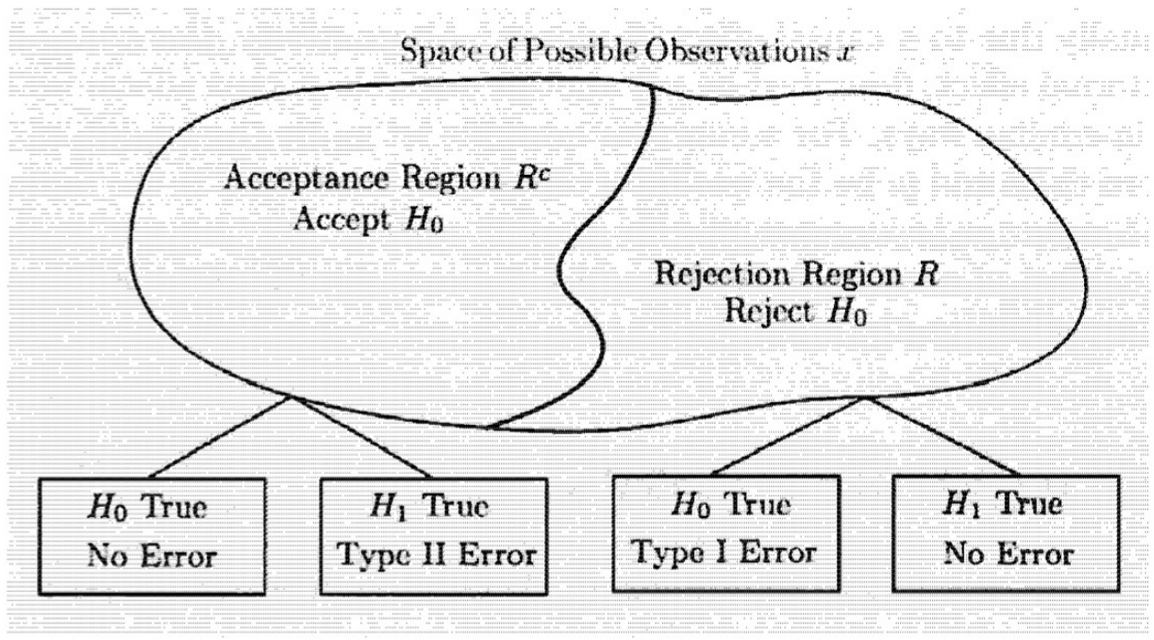
- Binary θ; new terminology:
  - **null hypothesis** H_0:
    $$X \sim p_X(x; H_0) \quad [\text{or } f_X(x; H_0)]$$
  - **alternative hypothesis** H_1:
    $$X \sim p_X(x; H_1) \quad [\text{or } f_X(x; H_1)]$$



Binary Hypothesis Testing

- Binary θ; new terminology:
  - **null hypothesis** H_0:
    $$X \sim p_X(x; H_0) \quad [\text{or } f_X(x; H_0)]$$
  - **alternative hypothesis** H_1:
    $$X \sim p_X(x; H_1) \quad [\text{or } f_X(x; H_1)]$$

- Partition the space of possible data vectors
  Rejection region R:
  reject H_0 iff data $\in R$



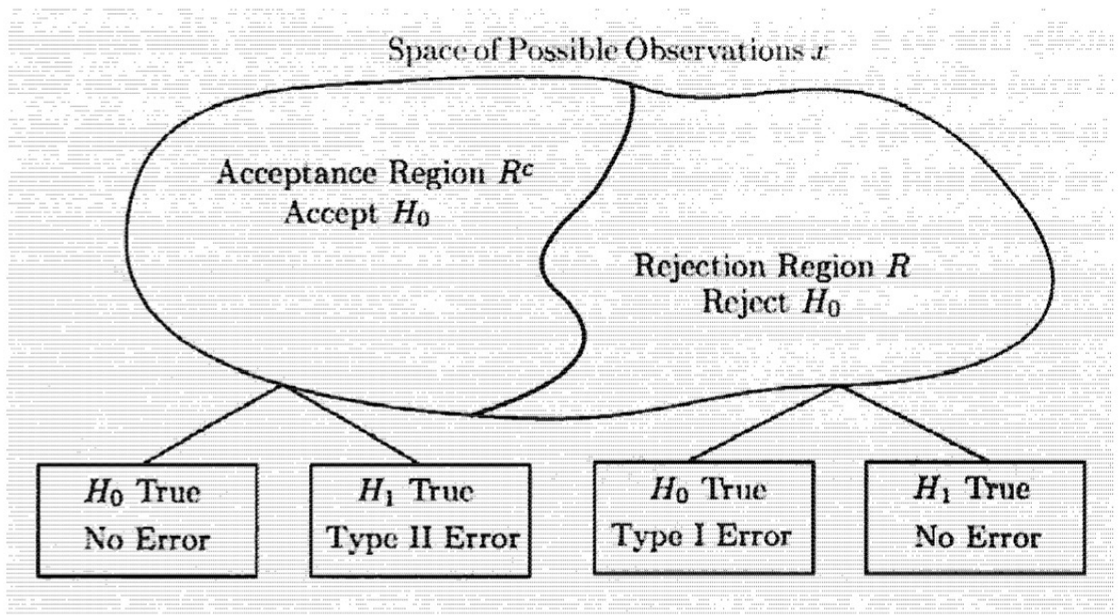
Binary Hypothesis Testing

- Partition the space of possible data vectors
  Rejection region R:
  reject H_0 iff data $\in R$

- Types of errors:
  - **Type I** (**false rejection**, false alarm):
    H_0 true, but rejected

$$\alpha(R) \;=\; \mathbf{P}(X \in R\,; H_0)$$

  - **Type II** (**false acceptance**, missed detection):
    H_0 false, but accepted



Bayesian Hypothesis Testing

- Parameter Θ takes one of m values $\{\theta_1, \ldots, \theta_m\}$.
- Hypothesis $H_i \equiv$ event $\{\Theta = \theta_i\}$
- Hypothesis Testing: given observation x, select one of the hypotheses $H_1, \ldots, H_m$
- **MAP Rule**: Select the hypothesis with the largest posterior probability
  - Select H_i if $\mathbb{P}(\Theta = \theta_i | X = x) = p_{\Theta|X}(\theta_i|x)$ is largest

$$\text{if } p_\Theta(\theta_i) p_{X|\Theta}(x|\theta_i) \text{ (if } X \text{ is discrete)}$$
$$\text{if } p_\Theta(\theta_i) f_{X|\Theta}(x|\theta_i) \text{ (if } X \text{ is continuous) is largest}$$

Example Ia

- We have two biased coins; coin 1 and coin 2; with biases (i.e. probability of heads) equal to p_1 and p_2 respectively. We choose a coin at random, and want to infer its identity based on the outcome of a single toss. Use the MAP Rule to do so.

Example Ia

- We have two biased coins; coin 1 and coin 2; with biases (i.e. probability of heads) equal to p_1 and p_2 respectively. We choose a coin at random, and want to infer its identity based on the outcome of a single toss. Use the MAP Rule to do so.

  - Consider random variables Θ, X.
    $\Theta = 1$ indicates coin 1, and $\Theta = 2$ indicates coin 2.
    $X = 1$ is coin flips heads, $X = 0$ for tails.

  - Suppose X = 0 (tails). Then, we select hypothesis Θ = 1 if

$$\begin{aligned} p_\Theta(1)p_{X|\Theta}(0|1) &> p_\Theta(2)p_{X|\Theta}(0|2) \\ (1-p_1) &> (1-p_2) \end{aligned}$$

Example Ib

- We have two biased coins; coin 1 and coin 2; with biases (i.e. probability of heads) equal to p_1 and p_2 respectively. We choose a coin at random, and want to infer its identity based on the outcome of **n** coin tosses. Use the MAP Rule to do so.

Example Ib

- We have two biased coins; coin 1 and coin 2; with biases (i.e. probability of heads) equal to p_1 and p_2 respectively. We choose a coin at random, and want to infer its identity based on the outcome of **n** coin tosses. Use the MAP Rule to do so.

  - Consider random variables Θ, X.
    $\Theta = 1$ indicates coin 1, and $\Theta = 2$ indicates coin 2.
    $X = k$ if k heads in the n coin flips.

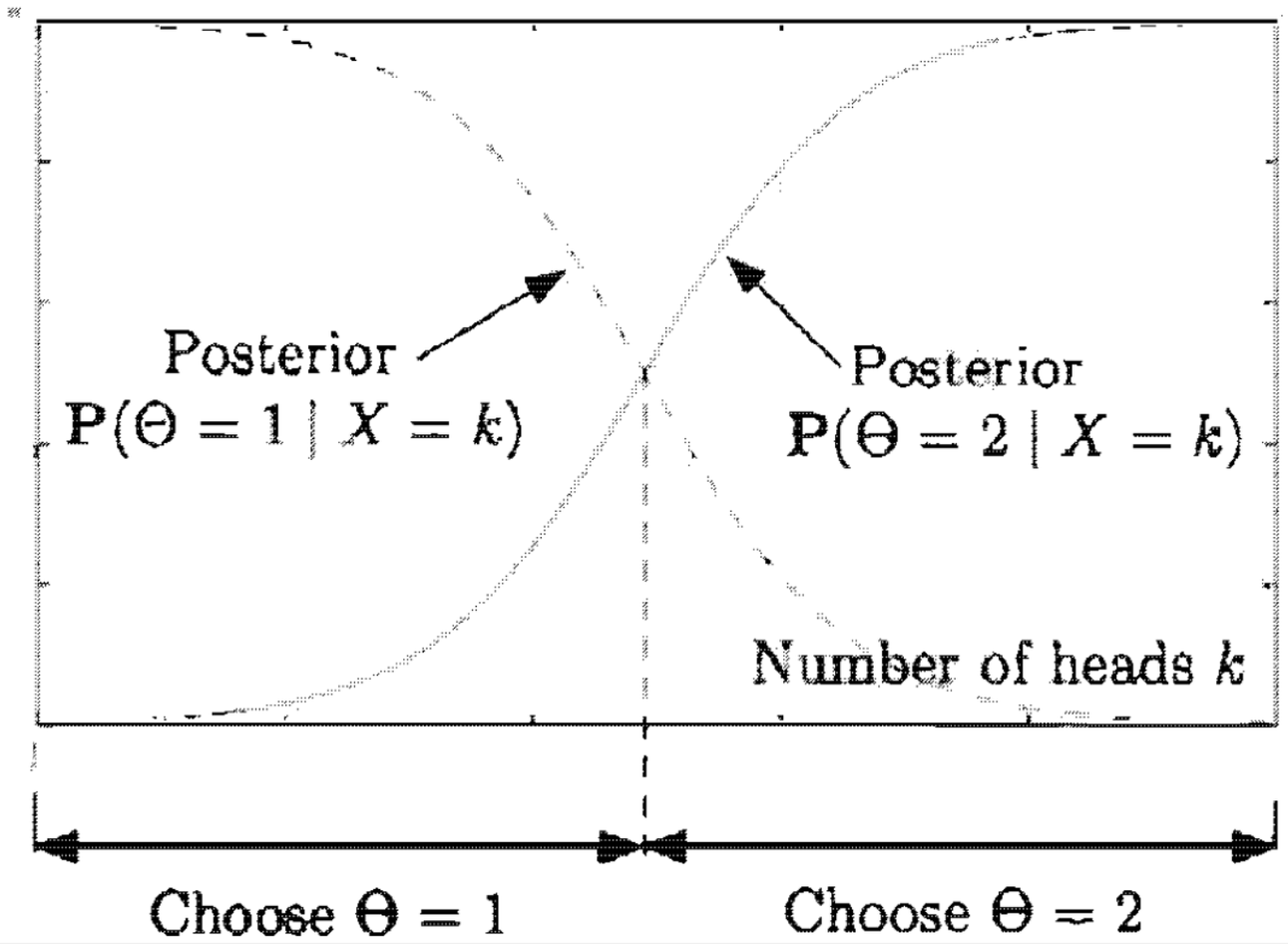
  - Suppose X = k. Then, we select hypothesis Θ = 1 if

$$\begin{aligned} p_\Theta(1)p_{X|\Theta}(k|1) &> p_\Theta(2)p_{X|\Theta}(k|2) \\ p_1^k(1-p_1)^{n-k} &> p_2^k(1-p_2)^{n-k} \end{aligned}$$

Example Ib

- Suppose X = k. Then, we select hypothesis Θ = 1 if

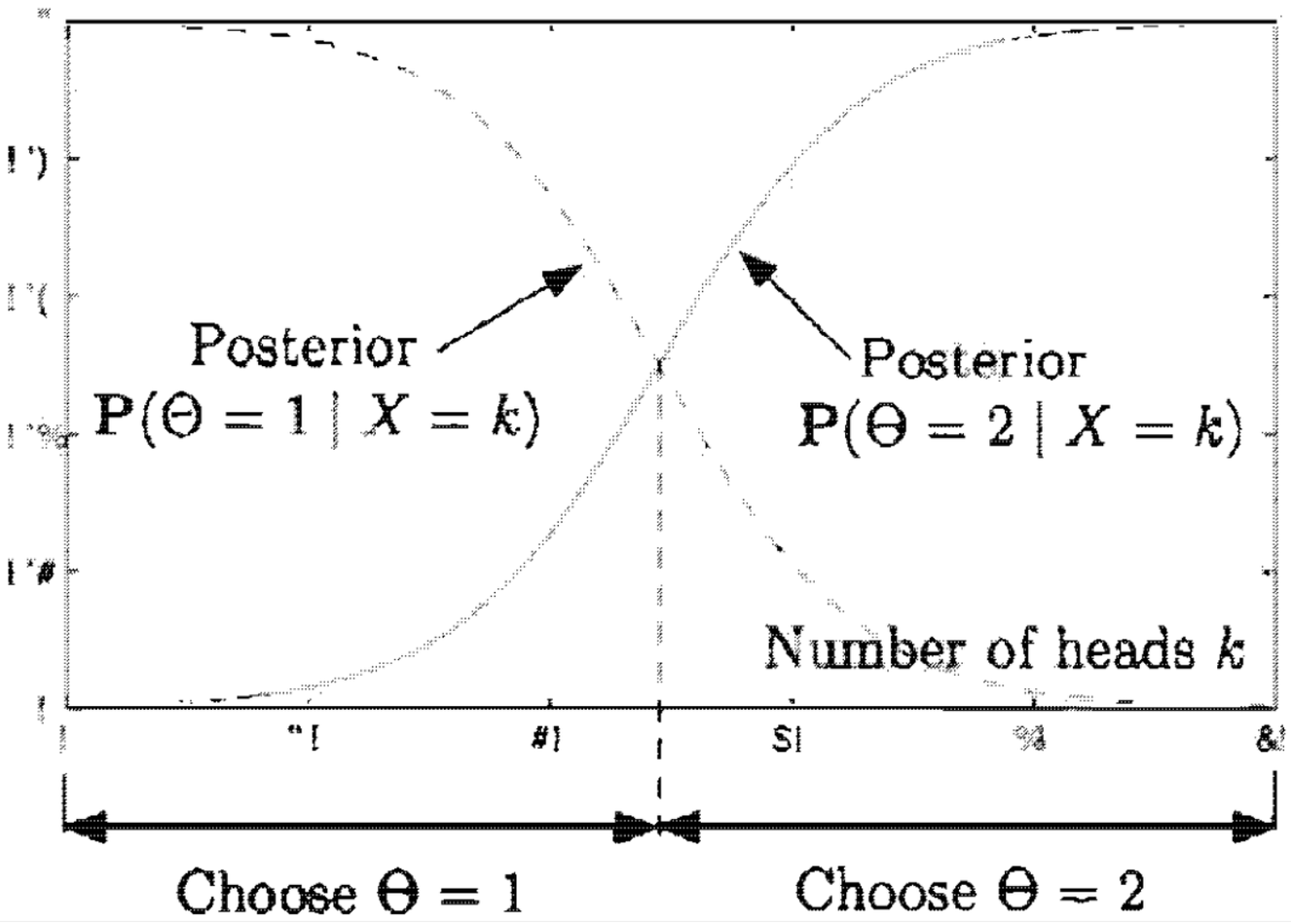
$$p_1^k(1-p_1)^{n-k} \quad > \quad p_2^k(1-p_2)^{n-k}$$



Example Ib

- Suppose X = k. Then, we select hypothesis \Theta = 1 if

$$p_1^k(1-p_1)^{n-k} \quad > \quad p_2^k(1-p_2)^{n-k}$$



We see that MAP Rule is quite simple:

$\Theta = 1$ iff $X \leq k^*$.