STAT 234 Lecture 18B
The General Framework of Hypothesis Testing
Section 9.1

Yibi Huang
Department of Statistics
University of Chicago

Can Dogs Smell Cancer?

Dogs Can Smell Cancer | Secret Life of Dogs | BBC

- https://youtu.be/e0UK6kkS0_M

Case Study: Can Dogs Smell Bladder Cancer?

- A study[1] by M. Willis et al. considered whether dogs could be trained to detect if a person has bladder cancer by smelling his/her urine.
- 6 dogs of varying breeds were trained to discriminate between urine from patients with bladder cancer and urine from control patients without it.
- The dogs were taught to indicate which among several specimens was from the bladder cancer patient by lying beside it.
- Once trained, the dogs' ability to distinguish cancer patients from controls was tested using urine samples from subjects not previously encountered by the dogs.

[1]Olfactory detection of human bladder cancer by dogs: proof of principle study, *British Medical Journal*, vol. 329, September 25, 2004.

Case Study: Can Dogs Smell Bladder Cancer?

- Neither the dog handlers nor the experimental observers knew the identity of urine samples so the dogs couldn't get clue
- Each of the 6 dogs was tested with 9 trials. In each trial, one urine sample from a bladder cancer patient was randomly placed among 6 control urine samples.
- Outcome: In the total of 54 trials with the 6 dogs, the dogs made the correct selection 22 times.
  - The dogs were correct for $22/54 \approx 41\%$ of the time,
    - not fabulous
  - If the dogs just guessed at random, they were only expected to be correct for $1/7 \approx 14\%$ of the time
  - Is this difference (41% v.s. 14%) surprising?

Two Competing Hypotheses

Let p be the probability that a dog makes the correct selection on a given trial.

- *Null hypothesis (H_0)*: $p = 1/7$
  "There is nothing going on."
  The dogs just guessed at random.
  - "null" means "nothing surprising is going on".
  - The dogs were just lucky to make more correct selections than expected.

Two Competing Hypotheses

Let p be the probability that a dog makes the correct selection on a given trial.

- *Null hypothesis (H_0)*: $p = 1/7$
  "There is nothing going on."
  The dogs just guessed at random.
  - "null" means "nothing surprising is going on".
  - The dogs were just lucky to make more correct selections than expected.

- *Alternative hypothesis (H_A or H_1)*: $p > 1/7$
  "There is something going on."
  Dogs can do better than random guessing.

Weighing Evidence Using a Test Statistic

The next step of hypothesis testing is to weigh the evidence —

how likely to observed the data obtained if H_0 was true?

- If the observed result was very unlikely to have occurred under the H_0, then the evidence raises more than a reasonable doubt in our minds about the H_0.

Weighing Evidence Using a Test Statistic

The next step of hypothesis testing is to weigh the evidence — how likely to observed the data obtained if H_0 was true?

- If the observed result was very unlikely to have occurred under the H_0, then the evidence raises more than a reasonable doubt in our minds about the H_0.

The *test statistic* is a summary of the data that best reflects the evidence for or against the hypotheses.

- For this study, the test statistics we choose is

$$X = \text{the number of correct guesses in the 54 trials}$$

- The larger X, the stronger evidence for H_A and against H_0
- The smaller X, the stronger evidence for H_0 and against H_A

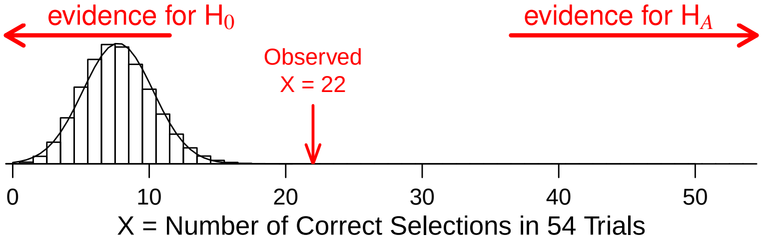
Distribution of the Test Statistics Under H_0

For the "Dogs Smell Cancer" study, if H_0 is true, then

$$X \sim Bin(n = 54,\ p = 1/7) \quad \text{(Why?)}$$

which implies

$$P(X = k) = \binom{54}{k}\left(\frac{1}{7}\right)^k \left(\frac{6}{7}\right)^{54-k}, \quad k = 0, 1, 2, \ldots, 54.$$



Test Procedure & Rejection Region

A test procedure is specified by the following:

1. a *test statistic*
2. a *rejection region*

The null hypothesis H_0 will be **rejected** if and only if **the test statistic falls in the rejection region**.

Test Procedure & Rejection Region

A test procedure is specified by the following:

1. a *test statistic*
2. a *rejection region*

The null hypothesis H_0 will be **rejected** if and only if **the test statistic falls in the rejection region**.

E.g., for the "Dogs Smell Cancer" study, as the strength of evidence for the two hypotheses are reflected by the test statistic

$$X = \text{\# of correct guesses in the 54 trials.}$$

A sensible **rejection region** is of the form

$$X \geq k \quad \text{for some cutoff } k.$$

and the test procedure is $\boxed{\text{reject } H_0 \text{ if } X \geq k}$.

Test Procedure & Rejection Region

A test procedure is specified by the following:

1. a *test statistic*
2. a *rejection region*

The null hypothesis H_0 will be **rejected** if and only if **the test statistic falls in the rejection region**.

E.g., for the "Dogs Smell Cancer" study, as the strength of evidence for the two hypotheses are reflected by the test statistic

$$X = \text{\# of correct guesses in the 54 trials.}$$

A sensible **rejection region** is of the form

$$X \geq k \quad \text{for some cutoff } k.$$

and the test procedure is $\boxed{\text{reject } H_0 \text{ if } X \geq k}$.

How to choose the cutoff value k for the rejection region?

Type I and Type II Errors

In a hypothesis test, we make a decision about which of H_0 or H_A might be true, but our decision might be incorrect.

Type I and Type II Errors

In a hypothesis test, we make a decision about which of H_0 or H_A might be true, but our decision might be incorrect.

		Decision	
		fail to reject H_0	reject H_0
Truth	H_0 true		
	H_A true		

Type I and Type II Errors

In a hypothesis test, we make a decision about which of H_0 or H_A might be true, but our decision might be incorrect.

		Decision	
		fail to reject H_0	reject H_0
Truth	H_0 true	✓	
	H_A true		

Type I and Type II Errors

In a hypothesis test, we make a decision about which of H_0 or H_A might be true, but our decision might be incorrect.

		Decision	
		fail to reject H_0	reject H_0
Truth	H_0 true	✓	
	H_A true		✓

Type I and Type II Errors

In a hypothesis test, we make a decision about which of H_0 or H_A might be true, but our decision might be incorrect.

		Decision	
		fail to reject H_0	reject H_0
Truth	H_0 true	✓	*Type I Error*
	H_A true		✓

- A *Type I Error* is rejecting the H_0 when it is true.

Type I and Type II Errors

In a hypothesis test, we make a decision about which of H_0 or H_A might be true, but our decision might be incorrect.

		Decision	
		fail to reject H_0	reject H_0
Truth	H_0 true	✓	*Type I Error*
	H_A true	*Type II Error*	✓

- A *Type I Error* is rejecting the H_0 when it is true.
- A *Type II Error* is failing to reject the H_0 when it is false.

Significance Level $\alpha = P(\text{Type I error})$

The *significance level* α of a test procedure is its probability to reject the null hypothesis H_0 when H_0 is true.

$$\alpha = P(\text{Type I error}) = P(\text{reject } \mathrm{H}_0 \mid H_0 \text{ is true})$$

Significance Level $\alpha = P(\text{Type I error})$

The *significance level* α of a test procedure is its probability to reject the null hypothesis H_0 when H_0 is true.

$$\alpha = P(\text{Type I error}) = P(\text{reject } H_0 \mid H_0 \text{ is true})$$

For the "Dog Smell Cancer" Study, if the test procedure is rejecting H_0 if $X \geq 15$, the significance level would be

$$\begin{aligned}\alpha = P(\text{Type I error}) &= P(H_0 \text{ is rejected when } H_0\ (p = 1/7) \text{ is true}) \\ &= P(X \geq 15 \text{ when } X \sim Bin(n = 54, p = 1/7)) \\ &= \sum_{k=15}^{54} \binom{54}{k} \left(\frac{1}{7}\right)^k \left(\frac{6}{7}\right)^{54-k} \approx 0.0073\end{aligned}$$

```
sum(dbinom(15:54, size = 54, p = 1/7))
[1] 0.007288514
```

Significance Level $\alpha = P$(Type I error)

The *significance level* α of a test procedure is its probability to reject the null hypothesis H_0 when H_0 is true.

$$\alpha = P(\text{Type I error}) = P(\text{reject } H_0 \mid H_0 \text{ is true})$$

For the "Dog Smell Cancer" Study, if the test procedure is rejecting H_0 if $X \geq 15$, the significance level would be

$$\begin{aligned}\alpha = P(\text{Type I error}) &= P(H_0 \text{ is rejected when } H_0\ (p = 1/7) \text{ is true}) \\ &= P(X \geq 15 \text{ when } X \sim Bin(n = 54, p = 1/7)) \\ &= \sum_{k=15}^{54} \binom{54}{k} \left(\frac{1}{7}\right)^k \left(\frac{6}{7}\right)^{54-k} \approx 0.0073\end{aligned}$$

```
sum(dbinom(15:54, size = 54, p = 1/7))
[1] 0.007288514
```

If we reject H_0 when $X \geq 15$, there is a chance of 0.0073 to falsely reject a correct H_0 (Type I error).

Example (Dogs Smell Cancer)

For the test procedure: rejecting H_0 when $X \geq k$, the chance of making a Type I error is

$$\begin{aligned}
P(\text{Type I error}) &= P(\text{H}_0 \text{ is rejected when H}_0\ (p = 1/7) \text{ is true}) \\
&= P(X \geq k \text{ when } X \sim Bin(n = 54, p = 1/7)) \\
&= \sum_{x=k}^{54} \binom{54}{k} \left(\frac{1}{7}\right)^x \left(\frac{6}{7}\right)^{54-x} \approx \begin{cases} 0.14 & \text{if } k = 11 \\ 0.076 & \text{if } k = 12 \\ 0.038 & \text{if } k = 13 \\ 0.017 & \text{if } k = 14 \\ 0.007 & \text{if } k = 15 \end{cases}
\end{aligned}$$

Setting Rejection Region Based on the Significance Level

For the dogs study,

$$P(\text{Type I error}) = \begin{cases} 0.14 & \text{if rejecting } \mathrm{H}_0 \text{ when } X \geq 11 \\ 0.076 & \text{if rejecting } \mathrm{H}_0 \text{ when } X \geq 12 \\ 0.038 & \text{if rejecting } \mathrm{H}_0 \text{ when } X \geq 13 \\ 0.017 & \text{if rejecting } \mathrm{H}_0 \text{ when } X \geq 14 \\ 0.007 & \text{if rejecting } \mathrm{H}_0 \text{ when } X \geq 15 \end{cases}$$

Setting Rejection Region Based on the Significance Level

For the dogs study,

$$P(\text{Type I error}) = \begin{cases} 0.14 & \text{if rejecting } H_0 \text{ when } X \geq 11 \\ 0.076 & \text{if rejecting } H_0 \text{ when } X \geq 12 \\ 0.038 & \text{if rejecting } H_0 \text{ when } X \geq 13 \\ 0.017 & \text{if rejecting } H_0 \text{ when } X \geq 14 \\ 0.007 & \text{if rejecting } H_0 \text{ when } X \geq 15 \end{cases}$$

To determine the cutoff value k for the rejection region $\{X \geq k\}$, we can first choose a *significance level* α , which is *the maximal P(Type I error) we can tolerate*, and then choose the cutoff value so that P(Type I error) does not exceeds the significance level α.

Setting Rejection Region Based on the Significance Level

For the dogs study,

$$P(\text{Type I error}) = \begin{cases} 0.14 & \text{if rejecting } \mathrm{H}_0 \text{ when } X \geq 11 \\ 0.076 & \text{if rejecting } \mathrm{H}_0 \text{ when } X \geq 12 \\ 0.038 & \text{if rejecting } \mathrm{H}_0 \text{ when } X \geq 13 \\ 0.017 & \text{if rejecting } \mathrm{H}_0 \text{ when } X \geq 14 \\ 0.007 & \text{if rejecting } \mathrm{H}_0 \text{ when } X \geq 15 \end{cases}$$

To determine the cutoff value k for the rejection region $\{X \geq k\}$, we can first choose a *significance level* α , which is *the maximal P(Type I error) we can tolerate*, and then choose the cutoff value so that P(Type I error) does not exceeds the significance level α.

- If we can tolerate a $\alpha = 5\%$ chance of Type I error, the test procedure can be "rejecting H_0 if $X \geq 13$"
- If we can tolerate a $\alpha = 1\%$ chance of Type I error, the test procedure can be "rejecting H_0 if $X \geq 15$"

A Smaller Significance Level Leads to a Higher P(Type II Error)

One might want to avoid a Type I error as much as possible by setting a tiny significance level. However,

smaller significance level $\Rightarrow$ smaller P(Type I error)

A Smaller Significance Level Leads to a Higher P(Type II Error)

One might want to avoid a Type I error as much as possible by setting a tiny significance level. However,

$$\begin{aligned}\text{smaller significance level} &\Rightarrow \text{smaller P(Type I error)} \\ &\Rightarrow \text{less likely to reject } H_0\end{aligned}$$

A Smaller Significance Level Leads to a Higher P(Type II Error)

One might want to avoid a Type I error as much as possible by setting a tiny significance level. However,

$$\begin{aligned} \text{smaller significance level} &\Rightarrow \text{smaller P(Type I error)} \\ &\Rightarrow \text{less likely to reject } H_0 \\ &\Rightarrow \text{more likely to make Type II error} \end{aligned}$$

A Smaller Significance Level Leads to a Higher P(Type II Error)

One might want to avoid a Type I error as much as possible by setting a tiny significance level. However,

$$\begin{aligned}\text{smaller significance level} &\Rightarrow \text{smaller P(Type I error)}\\ &\Rightarrow \text{less likely to reject } H_0\\ &\Rightarrow \text{more likely to make Type II error}\\ &\Rightarrow \text{higher P(Type II error)}\end{aligned}$$

A Smaller Significance Level Leads to a Higher P(Type II Error)

One might want to avoid a Type I error as much as possible by setting a tiny significance level. However,

$$\begin{aligned} \text{smaller significance level} &\Rightarrow \text{ smaller P(Type I error)} \\ &\Rightarrow \text{ less likely to reject } H_0 \\ &\Rightarrow \text{ more likely to make Type II error} \\ &\Rightarrow \text{ higher P(Type II error)} \end{aligned}$$

Suppose the sample size is fixed and a test statistic is chosen, choosing a rejection region with a smaller P(Type I error) would lead to a larger P(Type II error).

P-values

The *P-value* of test is the **probability of obtaining a test statistic such that the evidence for the alternative hypothesis H_A is *at least as strong* as our observed data, assuming the H_0 is true**.

P-values

The *P-value* of test is the **probability of obtaining a test statistic such that the evidence for the alternative hypothesis H_A is *at least as strong* as our observed data, assuming the H_0 is true**.

The definition is mouthful. Here are some key points

- The P-value is a *probability*, and thus it's between 0 and 1

P-values

The *P-value* of test is the **probability of obtaining a test statistic such that the evidence for the alternative hypothesis H_A is *at least as strong* as our observed data, assuming the H_0 is true**.

The definition is mouthful. Here are some key points

- The P-value is a *probability*, and thus it's between 0 and 1
- This probability is calculated *assuming the H_0 is true*.

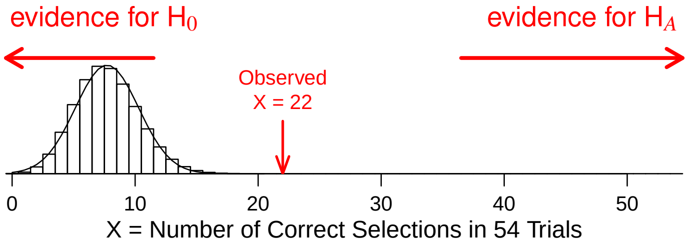
P-values

The *P-value* of test is the **probability of obtaining a test statistic such that the evidence for the alternative hypothesis H_A is *at least as strong* as our observed data, assuming the H_0 is true**.

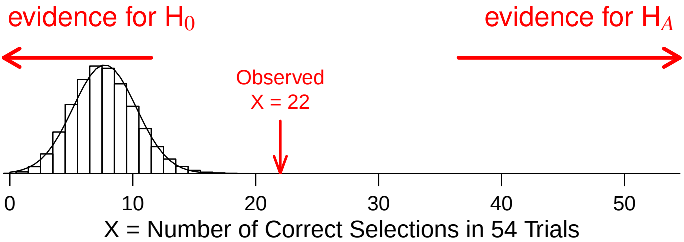
The definition is mouthful. Here are some key points

- The P-value is a *probability*, and thus it's between 0 and 1
- This probability is calculated *assuming the H_0 is true*.
- To determine the P-value, we must first decide which values of the test statistic are the evidence for H_A to be stronger than or as as the value obtained from our sample

Example (Dogs Smell Cancer) — P-Value



Example (Dogs Smell Cancer) — P-Value



- Observed $X = 22$
- Evidence for H_A is stronger than or as strong as the observed $X = 22$ if $X \geq 22$
- Under H_0, $X \sim Bin(n = 54,\ p = 1/7)$

$$P\text{-value} = P(X \geq 22 \mid H_0) = \sum_{k=22}^{54} \binom{54}{k} \left(\frac{1}{7}\right)^k \left(\frac{6}{7}\right)^{54-k} \approx 1.86 \times 10^{-6}$$

```
sum(dbinom(22:54,54,1/7))
[1] 1.861522e-06
```

P-Value as Strength of Evidence Against H_0

The smaller the P-value, the stronger the evidence against the H_0.

- A P-value of 0.25 says that if the H_0 was true, then we would obtain a result like the observed data 1 in 4 of the time; $\Rightarrow$ the data look consistent with H_0

P-Value as Strength of Evidence Against H_0

The smaller the P-value, the stronger the evidence against the H_0.

- A P-value of 0.25 says that if the H_0 was true, then we would obtain a result like the observed data 1 in 4 of the time; $\Rightarrow$ the data look consistent with H_0
- A P-value of 0.001 says that if the H_0 was true, then only 1 out of every 1,000 similar experiments would give result like the observed one; $\Rightarrow$ the H_0 looks doubtful

P-Value as Strength of Evidence Against H_0

The smaller the P-value, the stronger the evidence against the H_0.

- A P-value of 0.25 says that if the H_0 was true, then we would obtain a result like the observed data 1 in 4 of the time; $\Rightarrow$ the data look consistent with H_0
- A P-value of 0.001 says that if the H_0 was true, then only 1 out of every 1,000 similar experiments would give result like the observed one; $\Rightarrow$ the H_0 looks doubtful

P-Value as Strength of Evidence Against H_0

The smaller the P-value, the stronger the evidence against the H_0.

- A P-value of 0.25 says that if the H_0 was true, then we would obtain a result like the observed data 1 in 4 of the time; $\Rightarrow$ the data look consistent with H_0
- A P-value of 0.001 says that if the H_0 was true, then only 1 out of every 1,000 similar experiments would give result like the observed one; $\Rightarrow$ the H_0 looks doubtful

For the dogs study, if the dogs just guessed at random, there is less than 2 out of 1 million chance to be correct 22 or more times in 54 trials

- The observed result was very unlikely to have occurred under the H_0 — strong evidence to disbelieve H_0.

Test Procedure Based on the P-value

As an alternative to test procedures based on rejection regions, one can use test procedures based on P-values

1. Select a significance level α (as before, the desired P(type I error)).
2. Then
   - reject H_0 if the P-value $\leq \alpha$
   - do not reject H_0 if the P-value $> \alpha$

“Rejection Region” and “P-value” Approaches Are Equivalent

Using the Dogs study example, for a chosen significance level α, the rejection rejection $\{X \geq k\}$ must satisfy

$$P(X \geq k) \leq \alpha \quad \text{and} \quad P(X \geq k-1) > \alpha,$$

If the observed test statistic is $X = x_0$, the P-value would be

$$P\text{-value} = P(X \geq x_0)$$

“Rejection Region” and “P-value” Approaches Are Equivalent

Using the Dogs study example, for a chosen significance level α, the rejection rejection $\{X \geq k\}$ must satisfy

$$P(X \geq k) \leq \alpha \quad \text{and} \quad P(X \geq k-1) > \alpha,$$

If the observed test statistic is $X = x_0$, the P-value would be

$$P\text{-value} = P(X \geq x_0)$$

- If the observed $X = x_0$ falls in the rejection region $X \geq k$, then

  $$P\text{-value} = P(X \geq x_0) \leq P(X \geq k) \leq \alpha \quad \text{since } x_0 \geq k,$$

  then H_0 would be rejected by both test procedures.

"Rejection Region" and "P-value" Approaches Are Equivalent

Using the Dogs study example, for a chosen significance level α, the rejection rejection $\{X \geq k\}$ must satisfy

$$P(X \geq k) \leq \alpha \quad \text{and} \quad P(X \geq k-1) > \alpha,$$

If the observed test statistic is $X = x_0$, the P-value would be

$$P\text{-value} = P(X \geq x_0)$$

- If the observed $X = x_0$ falls in the rejection region $X \geq k$, then

  $$P\text{-value} = P(X \geq x_0) \leq P(X \geq k) \leq \alpha \quad \text{since } x_0 \geq k,$$

  then H_0 would be rejected by both test procedures.
- If the observed $X = x_0$ is NOT in the rejection region $X \geq k$, i.e., $x_0 \leq k - 1$, then

  $$P\text{-value} = P(X \geq x_0) \geq P(X \geq k-1) > \alpha \quad \text{since } x_0 \leq k-1,$$

  then H_0 would NOT be rejected by either approach.

P-value is the Smallest Significance Level to Reject H_0

The P-value is the **smallest significance level α at which the H_0 can be rejected**.

- e.g., the P-value for the dog study is 1.86×10^{-6}.
  The H_0 won't be rejected unless the significance level is as small as 1.86×10^{-6}

Because of this, the P-value is alternatively referred to as the *observed significance level* for the data.

Failing to Reject $H_0 \neq$ Accepting H_0

When the evidence is not strong enough to reject the H_0,
we say "we *fail to reject* the H_0" not "we *accept* the H_0"

- When we fail to reject the H_0, we might have made a Type II error

Failing to Reject $H_0 \neq$ Accepting H_0

When the evidence is not strong enough to reject the H_0,
we say "we *fail to reject* the H_0" not "we *accept* the H_0"

- When we fail to reject the H_0, we might have made a Type II error
- P(Type II error) can be quite high as it's not controlled.

Failing to Reject $H_0 \neq$ Accepting H_0

When the evidence is not strong enough to reject the H_0,
we say “we *fail to reject* the H_0” not “we *accept* the H_0”

- When we fail to reject the H_0, we might have made a Type II error
- P(Type II error) can be quite high as it’s not controlled.
- Recall so far we’ve only controlled P(Type I error) by the significance level but haven’t taken any measure to control P(Type II error)

True or False

If H_0 is rejected, then we can be certain that H_0 is false.

If H_0 is rejected at 5% level, there is less than a 5% chance for H_0 to be true.

True or False

If H_0 is rejected, then we can be certain that H_0 is false.

False. Even if H_0 is true, 5% of the time the experiment will give a result with a P-value $< 5\%$ so that H_0 is rejected.

If H_0 is rejected at 5% level, there is less than a 5% chance for H_0 to be true.

True or False

If H_0 is rejected, then we can be certain that H_0 is false.

False. Even if H_0 is true, 5% of the time the experiment will give a result with a P-value $< 5\%$ so that H_0 is rejected.

If H_0 is rejected at 5% level, there is less than a 5% chance for H_0 to be true.

False. A P-value does not give the chance of H_0 being true. In fact, the P-value is computed assuming H_0 is true.

$$P\text{-value} = P(data \mid H_0\ is\ true),\ \text{not}\ P(H_0\ is\ true \mid data).$$

Always Report the P-Value

Don't simply report the conclusion of whether H_0 is rejected. Always report the P-value

- A P-value of 0.04 and a P-value of 0.000001 are not at all the same thing, even though H_0 will be rejected at 0.05 level in both cases, but the strength of evidence are very different
- Simply reporting whether H_0 is rejected without P-value is like reporting the temperature as "cold" or "hot"
- It's much better to report the P-value and let people choose their own significance level, just like telling someone the temperature and let them decide for themselves whether they want to wear a coat

Conclusion of the Dogs Smell Bladder Cancer Study

- There is strong evidence that dogs have some ability to smell bladder cancer,
- However, the dogs were only correct 40% of the time, too low for practical application
- Another study (M. McCulloch et al., Integrative Cancer Therapies, vol 5, p. 30, 2006.) considered whether dogs could be trained to detect whether a person has lung cancer by smelling the subjects' breath. In one test with 83 Stage I lung cancer samples, the dogs correctly identified the cancer sample 81 times.

Recap: Hypothesis Testing Framework

1. We start with a *null hypothesis (H_0)* that represents the status quo.
2. We also have an *alternative hypothesis (H_A)* that represents our research question, i.e. what we're testing for.
3. We then collect data and often summarize the data as a *test statistic*, which is usually a measure gauging whether H_0 or H_A are more plausible
4. We then predict what the *test statistic* would be around under the assumption that the H_0 is true.
5. If the *test statistic* is too far away from what the H_0 predicts, we then reject the H_0 in favor of the H_A.

Recap: Hypothesis Testing Framework (Cont'd)

Using the "Rejection Region" Approach,

6. we choose a *significance level* α = maximal P(Type I error) that we can tolerate
7. we select the rejection region based on the significance level
8. we reject H_0 if the test statistic falls in the rejection region, and do not reject otherwise

Using the "P-value" Approach,

6. we calculate the P-value based on the test statistic
7. (optional) we choose *significance level* α = maximal P(Type I error) that we can tolerate and reject H_0 if the P-value $\leq \alpha$ and not to reject otherwise.

This lecture just introduces the general framework of hypotheses testing.

In the next several lectures, we will introduce several hypotheses tests for various types of problems.