Biost 517 / Biost 514 Applied Biostatistics I / Biostatistics I

Timothy A. Thornton, Ph.D.
Associate Professor of Biostatistics
University of Washington

Lecture 6:

Introduction to Hypothesis Testing;

Correct Interpretation of P-values;

One-sample t-test with R;

Tests of Significance

Outline

- General Procedure for Hypothesis Testing
  - Null and Alternative Hypotheses
  - Test Statistics
  - *p*-values
- Interpretation of the Significance Level
- Tests for a Population Mean
- Interpretation of *p*-values
- Statistical vs. Practical Significance
- Relationship between Confidence Intervals and Hypothesis Tests
- Potential Abuses of Tests

Types of inference

- Point estimation: what is the most likely value for this parameter in the population?
- Interval estimation: a range of values for a parameter that is likely to contain the true population value
- **Hypothesis testing: Are the data compatible with the true population value being some particular value?**
  - **Important that the particular value is stated prior to collecting/analyzing the data.**

Testing Hypotheses

- A confidence interval is a very useful statistical inference tool when the goal is to estimate a population parameter.
- When the goal is to assess the evidence provided by the data in favor of some claim about the population, **hypothesis tests**, or **tests of significance** are used.
- Common goal of **hypothesis testing**: Assess if the data are compatible with the true population value being some particular value
- A **hypothesis test** is an assessment of the evidence provided by the data in favor of (or against) some claim about the population.

Testing Hypotheses

- Suppose we perform a randomized experiment or take a random sample and calculate some sample statistic, say the sample mean.
- We want to decide if the ***observed*** value of the *sample* statistic is consistent with some *hypothesized* value of the corresponding parameter in the *population*.
- If the observed value in the sample and hypothesized value for the population differ (as they almost certainly will), is the difference due to an incorrect hypothesis or merely due to chance variation?

Components of a Hypothesis Test

- The **Null Hypothesis,** usually denoted H_0, is the statement being tested. Usually it states that the difference between the observed value from a sample and the hypothesized value is only due to chance variation.
  - Often the status quo
  - We will tend to believe this unless we "prove" otherwise – Often what we hope to disprove
- The **Alternative Hypothesis,** usually denoted H_A or H_1, is the statement that will be supported if the evidence suggests that the null hypothesis is false. It usually states that there is a real difference between the observed and hypothesized values.
  - Often what we hope is true

Hypothesis Testing: Aspirin Example

- **"Aspirin cuts cancer risk"**
- According to the American Cancer Society, the lifetime risk of developing colon cancer is 1 in 16. Taking an aspirin every other day for 20 years can cut your risk of colon cancer nearly in half, a study suggests. However, the benefits may not kick in until at least a decade of use
  - **SOURCE: The Associated Press, September 7, 1995**

Hypothesis Testing: Aspirin Example

- H_0: When taking aspirin every day, the lifetime risk of colon cancer is 1/16

- H_A: When taking aspirin every day, the lifetime risk of colon cancer is less than 1/16

Hypothesis Testing

- Null and alternative hypotheses are stated in terms of population parameters.
- In the previous example, let θ represent the lifetime risk of colon cancer in the population.
- Then
  - Null Hypothesis is H_0: $\theta = 1/16$ when taking aspirin everyday
  - Alternative Hypothesis is H_A: $\theta < 1/16$ when taking aspirin everyday
- In this example, the alternative hypothesis is “one-sided”

One and Two-sided alternatives

- Consider H_0: $\theta = 1/16$ when taking daily aspirin
- Some possible alternatives:
  - H_A: $\theta \neq 1/16$; a two-sided alternative
  - H_A: $\theta < 1/16$; one-sided alternative
  - H_A: $\theta > 1/16$; one-sided alternative
- The investigators, not the data, determine the appropriate alternative hypothesis. For research integrity, this choice should be made before the data are collected.

Hypothesis Testing Paradigm

In statistical hypothesis testing, the null hypothesis is "put to the test." Statisticians are trained to be *skeptics!*

In the Big Picture, the procedure is:

1. First assume the null hypothesis IS true
2. We then ask: Are the data unusual if the null hypothesis is true?
    - If yes, the decision is: "reject the null hypothesis"
    - If no, the decision is: "do not reject the null hypothesis"

Hypothesis Testing Paradigm

Details on how this is done in practice:

1. We use a “test statistic”. Ideally, we know the distribution of the test statistics when the null hypothesis is true
2. We compare the value of the test statistic in our data with the sampling distribution of the statistic when the null hypothesis is true.
    - If our test statistic value is unusual for that distribution, we reject the null hypothesis
    - If our test statistic value is not unusual for that distribution, we do not reject the null hypothesis

Hypothesis Testing Procedure

More details:

1. We use a "test statistic." Ideally, we know the **sampling distribution** of the test statistic when the null hypothesis is true.
2. We compare the value of the test statistic in our data with the sampling distribution of the statistic when the null hypothesis is true.
   - If our value is unusual for that distribution, we reject the null hypothesis
   - If our value is not unusual for that distribution, we do not reject the null hypothesis

Test Statistic for Hypothesis Testing

- For hypothesis testing, we must calculate the **test statistic** on which the test will be based.
- The test statistic usually measures the difference between the observed data and what would be expected *if* the null hypothesis were true.
- When H_0 is true, we expect the estimate based on the sample to take a value near the parameter value specified by H_0.
- Our goal is to answer the question, “How extreme is the value calculated from the sample from what we would expect under the null hypothesis?"

Test Statistic for Hypothesis Testing

- In many common situations the test statistic for a hypothesis test has the form

$$\frac{\text{estimate - hypothesized value}}{\text{standard deviation of the estimate}}$$

Example: Filling Prescription Bottles

- A machine at a pharmaceutical production plant is designed to fill 16 oz prescription bottles with various pharmaceuticals
- The actual amount varies slightly from bottle to bottle. From past experience, it is known that the SD is $\sigma = 0.2$ oz. A simple random sample of 100 bottles filled by the machine has a mean of $\bar{x} = 15.94$ oz per bottle.
  - Is this evidence that the machine needs to be recalibrated, or could this difference be a result of random variation?
- What is the null hypothesis?
  - H_0: The machine is properly calibrated with a mean of $\mu = 16$ oz
- What are possible alternative hypotheses H_A?

$$H_A : \mu \neq 16$$

$$H_A : \mu > 16$$

$$H_A : \mu < 16$$

Example: Filling Prescription Bottles

- For the prescription bottle example, we have that the mean of the sample of size 100 is $\bar{x} = 15.94$ oz. The population mean μ specified by the null hypothesis is μ =16 oz, which may also be written as $\mu = \mu_0$, where μ_0= 16 oz to indicate that hypothesised parameter value is under the null hypothesis H_0.
- The population SD "known" to be σ =0.2 oz
- A test statistic for testing the null hypothesis of μ_0 =16 oz is

$$z = \frac{15.94 - 16}{0.2 / \sqrt{100}} = -3$$

- We'll have more to say about how to evaluate this statistic in a moment

P-value of a test statistics

Once we have a test statistic value, we must obtain a *p*-value for the test statistic:

- The p-value is the probability of observing a test statistic *as extreme or more extreme than actually observed*, assuming the null hypothesis H_0 is true.
- The smaller the p-value, the stronger the evidence *against* the null hypothesis.
- if the p-value is as small or smaller than some number α (e.g. 0.01, 0.05), we say that the result is **statistically significant** at level α.
- α is called the **significance level** of the test.

P-value of a test statistics

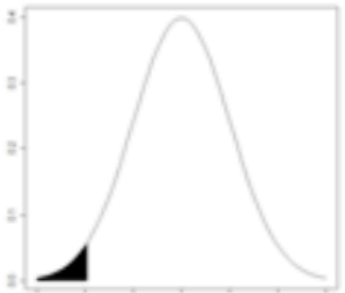
Once we have a test statistic value, we must obtain a *p*-value for the test statistic:

- test statistic → p-value
  - The key feature of a test statistic T is that the sampling distribution of T is known when H_0 is true.

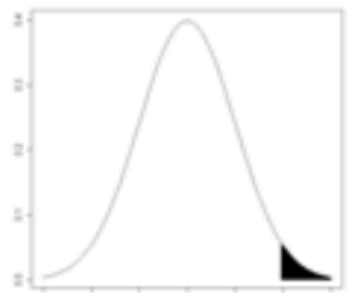
  Let $H_0 : \theta = \theta_0$ and t be the observed value of the test statistic.
  - Alternative Hypothesis:

$H_1 : \theta < \theta_0$ $\quad$ $H_1 : \theta > \theta_0$ $\quad$ $H_1 : \theta \neq \theta_0$

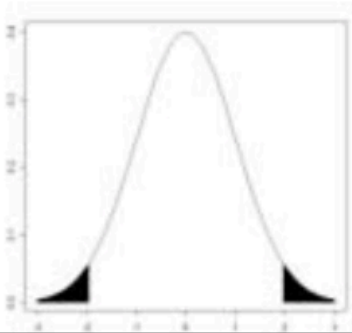
  - For 1-sided alternative hypothesis, the p-value is a single tail of the distribution of T under the null. For the 2-sided alternative, the p-value is doubled.



$p = P(T \leq t \mid \theta_0)$



$p = P(T \geq t \mid \theta_0)$



$p = 2 \times \min\left(P(T \leq t \mid \theta_0), P(T \geq t \mid \theta_0)\right)$

P-value of a test statistics

Suppose we want to test the hypothesis that μ has a specific value:

$$H_0 : \mu = \mu_0$$

Since $\bar{x}$ estimates μ, the test is based on $\bar{x}$, which has a (perhaps approximately) Normal distribution. Thus,

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

is a standard normal random variable, *under the null hypothesis.*

p-values for different alternative hypotheses:

- $H_a : \mu > \mu_0$ – p-value is $P(Z \geq z)$ (area of right-hand tail)
- $H_a : \mu < \mu_0$ – p-value is $P(Z \leq z)$ (area of left-hand tail)
- $H_a : \mu \neq \mu_0$ – p-value is $2P(Z \geq |z|)$ (area of both tails)

Hypothesis Testing Procedure

By transforming a test statistic to a p-value we can interpret the strength of the evidence against the null hypothesis on the probability scale.

$$0 \leq \text{p-value} \leq 1$$

We compare the p-value to a **pre-selected** value **α**. A common choice is α=0.05

- If p-value < α, we reject the null hypothesis
- If p-value ≥ α, we do not reject the null hypothesis.

Hypothesis Testing: Recap

Before analyzing the data:

- Specify null hypothesis H_0.
- Specify alternative hypothesis H_A.
- Specify the significance level α.

Data analysis:

- Compute test statistic T on the data.
- Compare observed value of T to the sampling distribution of T when the null hypothesis H_0 is true
- Use the alternative hypothesis to get the appropriate p-value P (one-sided or two-sided).
- Compare P to α and reject the null hypothesis if $P < \alpha$.

Interpretation of Significance Level

To perform a **test of significance level** α, we perform the previous three steps and then *reject* H_0 *if the p-value is less than* α.

The following outcomes are possible when conducting a test:

Reality	Our Decision	
	H_0	H_a
H_0	$\surd$	Type I Error
H_a	Type II Error	$\surd$

Suppose H_0 is actually true. If we draw many samples, and perform a test for each one, α of these tests will (incorrectly) reject H_0. In other words, α *is the probability that we will make a Type I error.*

Type II error is related to the notion of the *power* of a test, which we will discuss later.

Type I and Type II errors

Decision based on the data	Truth about Population: Null Hypothesis is True	Truth about Population: Alternative Hypothesis is True
Do not reject the null hypothesis	Correct Decision	Type II Error
Reject the null hypothesis	Type I error	Correct Decision

Quantifying Errors

- P(type I error | H_0) = α
  - "alpha-level" of the test
  - "significance level" of the test
  - "size" of the test
- P(type II error | H_1) = β
  - P(correct decision | H_1) = 1– β. 1– β is called the "power" of the test

- α is chosen before data analysis. A traditional choice is α=0.05 (many think this is too lenient and should be smaller).

Quantifying Errors

- Hypothesis tests are designed to limit chances of type I errors.
  - P(type I error | H_0) = α
- Often, null hypotheses are simple and alternative hypotheses are composite. For example,
  - H_0: $\mu = \mu_0$ vs Alt H_1: $\mu \neq \mu_0$

– Therefore, β will not be a constant. Instead, β will depend on the specific true value of the parameter
  - $\beta \equiv$ P(type II error | H_1) = P(type II error | $\mu = \mu_1 \neq \mu_0$)

– The farther the truth is from the Null value, the lower β
  - In other words, it is easier to detect big differences than small differences.

One- vs Two-sided Tests

- One sided test of greater alternative
  - Null H_0: $\mu = \mu_0$ (or Null H_0: $\mu \leq \mu_0$) vs Alt H_1: $\mu > \mu_0$
- One sided test of lesser alternative
  - Null H_0: $\mu = \mu_0$ (or Null H_0: $\mu \geq \mu_0$) vs Alt H_1: $\mu < \mu_0$
- Two sided test
  - Null H_0: $\mu = \mu_0$ vs Alt H_1: $\mu \neq \mu_0$

One- vs Two-sided

- Choose to fit your situation

- E.g., New treatment vs Placebo
  - One-sided test
    - Only approve new treatment if better

- E.g., Two existing treatments A and B
  - Two-sided test
    - Push treatment A if better
    - Neutral if about equal
    - Push treatment B if better

One- vs Two-sided

- Guidance: If your scientific theory specifies a direction of effect, consider a one-sided alternative. Otherwise, default to a two-sided test.
- The choice sometimes comes down to what we care about.
- A company packaging ground coffee sold in 1 pound bags may do quality control to check that the factory is shipping bags with an average weight of 1 pound (and small variability from bag to bag).
  - Two-sided alternative
- A consumer watch-dog group checks on the bags of coffee in the market to make sure consumers are not being cheated. They care about bags with too little coffee, so
  - H_0: $\mu \geq 1.0$ pound
  - H_A: $\mu < 1.0$ pound (one-sided alternative)

Prescription Bottles Example: Continued

We are interested in assessing whether or not the machine needs to be recalibrated, which will be the case if it is systematically over- or under-filling bottles. Thus, we will use the hypotheses

$$H_0 : \mu = 16$$
$$H_a : \mu \neq 16$$

Recall that $\bar{x} = 15.94$, $\sigma = 0.2$, and $n = 100$.
Thus,

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = -3$$

The p-value for a two-sided test is
$p = 2P(Z \geq 3) = 0.0026$.

If $\alpha = 0.01$, we reject H_0.

If $\alpha = 0.05$, we reject H_0.

What would the p-value be if the alternative hypothesis was H_A: $\mu < 16$?

Example 2: TV Tubes

TV tubes are taken at random and the lifetime measured. $n = 100, \sigma = 300$ and $\bar{x} = 1265$ (days). Test whether the population mean is 1200, or greater than 1200.

$$H_0 : \mu = 1200$$
$$H_a : \mu > 1200$$

Under $H_0, \bar{x} \sim N(1200, 30)$.

$\therefore z = \frac{\bar{x}-1200}{30} \sim N(0,1)$ under H_0

The test statistic is $z = \frac{1265-1200}{30} = 2.17$, and the p-value is $P(Z \geq 2.17 | H_0) = 0.015$

This is evidence against H_0 at significance level 0.05, so we reject H_0. That is, we conclude that the average lifetime of TV tubes is greater than 1200 days.

Population Mean Inference: Hypothesis Testing with Unknown σ

We have already seen that when we take a SRS, $x_1, x_2, \ldots, x_n$, from a population with unknown μ and known σ, if either

- the population is Normally distributed, or
- n is large enough,

then

$$\bar{X} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

Equivalently, the standardized sample mean, or the one-sample z statistic is

$$z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$$

What if the population SD σ is unknown? (A far more likely occurrence.)

Student's *t* Distribution

If the true population SD, σ, is unknown, we estimate σ using the sample SD S.

When the standard deviation of a statistic is estimated from the data, the result is called the standard error of the statistic. The standard error of the sample mean $\bar{x}$ is $SE_{\bar{X}} = \frac{S}{\sqrt{n}}$.

Now, instead of dealing with

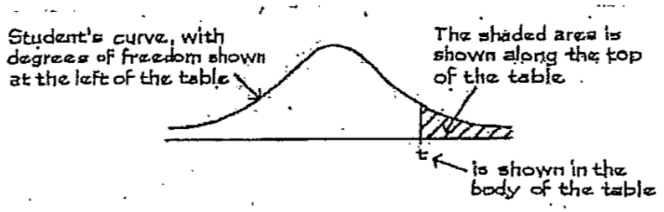
$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

we are interested in the quantity

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{(n-1)}$$

Here, $t_{(n-1)}$ is Student's *t* distribution, with $n - 1$ *degrees of freedom.* A table for the *t* distribution is given on the next page

Student's *t* Distribution

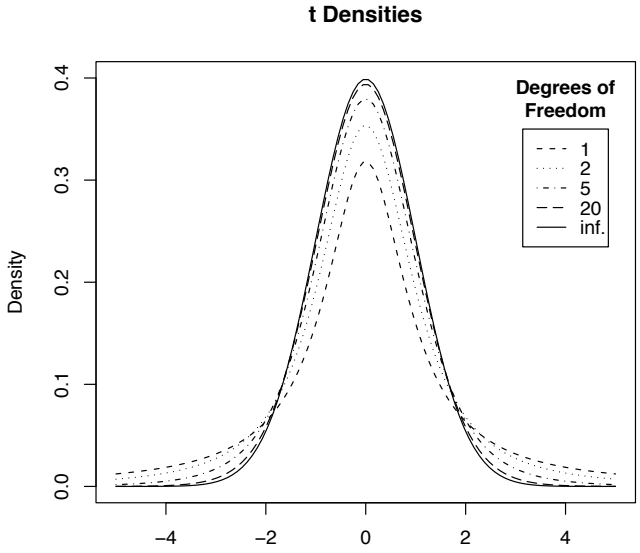


Degrees of freedom	*25%*	*10%*	*5%*	*2.5%*	*1%*	*0.5%*
1	1.00	3.08	6.31	12.71	31.82	63.66
2	0.82	1.89	2.92	4.30	6.96	9.92
3	0.76	1.64	2.35	3.18	4.54	5.84
4	0.74	1.53	2.13	2.78	3.75	4.60
5	0.73	1.48	2.02	2.57	3.36	4.03
6	0.72	1.44	1.94	2.45	3.14	3.71
7	0.71	1.41	1.89	2.36	3.00	3.50
8	0.71	1.40	1.86	2.31	2.90	3.36
9	0.70	1.38	1.83	2.26	2.82	3.25
10	0.70	1.37	1.81	2.23	2.76	3.17
11	0.70	1.36	1.80	2.20	2.72	3.11
12	0.70	1.36	1.78	2.18	2.68	3.05
13	0.69	1.35	1.77	2.16	2.65	3.01
14	0.69	1.35	1.76	2.14	2.62	2.98
15	0.69	1.34	1.75	2.13	2.60	2.95
16	0.69	1.34	1.75	2.12	2.58	2.92
17	0.69	1.33	1.74	2.11	2.57	2.90
18	0.69	1.33	1.73	2.10	2.55	2.88
19	0.69	1.33	1.73	2.09	2.54	2.86
20	0.69	1.33	1.72	2.09	2.53	2.85
21	0.69	1.32	1.72	2.08	2.52	2.83
22	0.69	1.32	1.72	2.07	2.51	2.82
23	0.69	1.32	1.71	2.07	2.50	2.81
24	0.68	1.32	1.71	2.06	2.49	2.80
25	0.68	1.32	1.71	2.06	2.49	2.79

Reminder: Properties of Student's *t* Distribution

- Symmetric about zero
- Bell-shaped, similar to normal distribution
- More spread out than normal, i.e., heavier tails than a normal
- Exact shape depends on the degrees of freedom
- As the number of degrees of freedom increases, the *t* distribution converges to the Normal distribution.

The Density of Student's t



One Sample *t* test

Suppose a SRS of size n is drawn from a $N(\mu, \sigma)$ population with both μ and σ unknown. The t-statistic,

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}}$$

has the t distribution with $n - 1$ d.f.

To test $H_0 : \mu = \mu_0$, compute the one-sample t statistic,

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

The p-values are

$$\begin{aligned} H_a : \mu > \mu_0 \quad & P(t_{n-1} \geq t) \\ H_a : \mu < \mu_0 \quad & P(t_{n-1} \leq t) \\ H_a : \mu \neq \mu_0 \quad & 2P(t_{n-1} \geq |t|) \end{aligned}$$

These are exact if the population is normal, and otherwise approximately correct for large n.

Hypothesis Tests of Means

- One sample t-test: To test hypotheses that the population mean is some particular value

- Null Hypothesis
  - Usually status quo
    - We will tend to believe this unless we “prove” otherwise
  - Often what we hope to disprove

- Alternative Hypothesis
  - Often what we hope is true

Hypothesis Testing with *t* distribution: Example

Let X (in mm) denote the growth in 15 days of a tumor induced in a mouse. It is known from a previous experiment that the average tumor growth is $4mm$. A sample of 20 mice that have a genetic variant hypothesized to be involved in tumor growth yielded $\bar{x} = 3.8mm, s = 0.3mm$. Test whether $\mu = 4$ or not, assuming growths are normally distributed.

1. State the hypotheses

$$H_0 : \mu = 4 \qquad H_a : \mu \neq 4$$

2. Calculate the t-statistic

$$t = \frac{3.8 - 4.0}{0.3/\sqrt{20}} = -2.98$$

3. Determine the p-value using the t-distribution table:

$$p = 2\mathrm{P}(t_{19} \geq 2.98) < 2\mathrm{P}(t_{19} \geq 2.861) = 2(.005) = .01$$

Since p is less than 0.01, we reject H_0 at significance level $\alpha = 0.01$. (p-value=0.008)

R: CDF for t-distribution

- The ***pt()*** function in R can be used to obtain values for the cumulative distribution function for a t distribution. Use the help function in R or more details: ***?pt***
- For example, to obtain the probability that a random variable T has a value less than 2, P(T≤2), where T has a *t* distribution with 30 degrees of freedom, the following command can be used:

```
> pt(2,df=30)
[1] 0.9726875
```

Example: P-value in R with t distribution

- In the previous example for tumor growth on mice on page 37, we to calculate the p-value using a t-distribution with 19 degrees of freedom. This p-value is 2 times the probability of observing a value at least as large as 2.98 for a t distribution with 19 degrees of freedom. So we need to obtain 1 minus the cdf since

$$P(t_{19} \geq 2.98) = 1 - P(t_{19} < 2.98)$$

- We calculate the p-value in R as follows:

```
> 2*(1-pt(2.98,df=19))
[1] 0.007694802
```

- So the p-value is p=0.0077

A Rough Interpretation of *P*-values

p-value	Interpretation
$p > 0.10$	no evidence against H_0
$0.05 < p \leq 0.10$	weak evidence against H_0
$0.01 < p \leq 0.05$	evidence against H_0
$p \leq 0.01$	strong evidence against H_0

Statistical vs. Practical Significance

Saying that a result is *statistically significant* does not signify that it is large or necessarily important. That decision depends on the particulars of the problem. A statistically significant result only says that there is substantial evidence that H_0 is false.

Failure to reject H_0 does not imply that H_0 is correct. It only implies that *we have insufficient evidence to conclude that* H_0 *is incorrect.*

R: One-sample t-test

- Can perform a one-sample t-test using the **t.test()** function in R quite easily.
- Consider a sample that with with values of a variable of interest stored as a vector object named *var1* in R. To perform a one sample t-test for the population mean for *var1* being equal a fixed *nullval, the following commands can be used:*

```
t.test(var1,mu = nullval,alternative="two.sided")
```

- The output provides
  - p-values for a test that the mean is equal to *nullval*
  - provide 95% confidence intervals
- Can also do a one-sided tests with the alternative option by changing "two.sided" in the above command twith either "less" or "greater"

Example: one-sample t-test

- Total cholesterol levels less than or equal to 200 mg/dl are considered to be in the healthy range. Scientific question of interest: do diabetic men under a doctor's care have mean cholesterol levels that are in the healthy range? In a study of Cardiovascular Health Study of elderly adults aged 65 years, 382 men diagnosed with diabetes at the time of enrollment. Perform a statistical test to assess if the data provide evidence that elderly diabetic men have a different mean cholesterol level than the highest level to be considered healthy, e.g., 200 mg/dl ?
- In this sample the two Hypotheses:

H_0: $\mu = 200$ mg/dl

H_a: $\mu \neq 200$ mg/dl

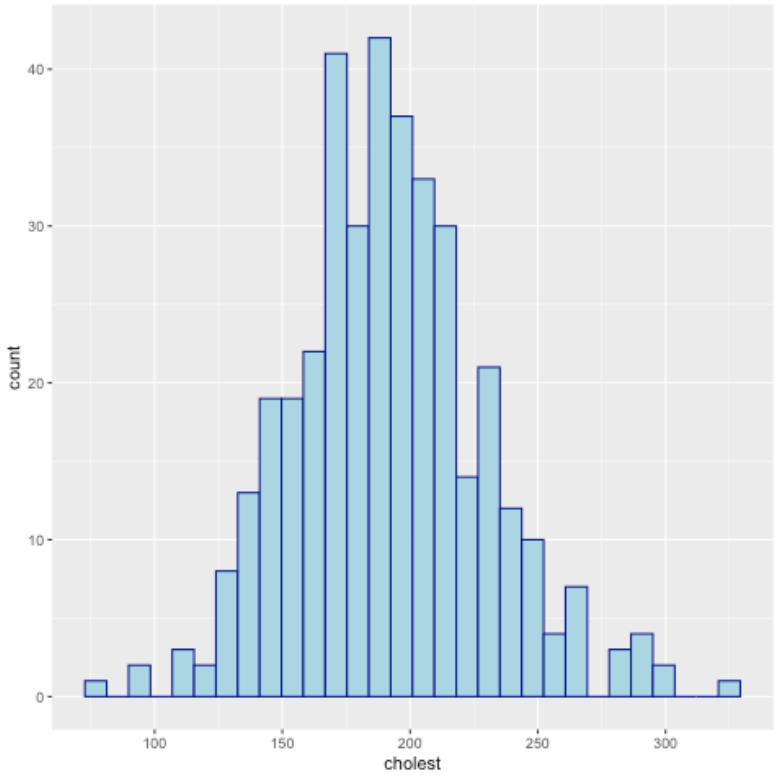
Example: one-sample t-test

- Total cholesterol levels less than or equal to 200 mg/dl are considered to be in the healthy range. Scientific question of interest: do diabetic men have mean cholesterol levels that are in the healthy range? In the Cardiovascular Health Study of elderly adults aged 65 years, there were 380 men diagnosed with diabetes at the time of enrollment with cholesterol measurements at baseline.
- Statistical question: assess if the data provide evidence that elderly diabetic men have a different mean cholesterol level than the highest level to be considered healthy, e.g., 200 mg/dl ?
- In this sample the two Hypotheses:

H_0: $\mu = 200$ mg/dl

H_a: $\mu \neq 200$ mg/dl

Example: Cholesterol Levels in Male Diabetics




```
> with(inflamm, summary(cholest[diab2==1&male==1]))
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
   73.0   168.8   189.0   191.5   213.0   321.0      22
> with(inflamm, sd(cholest[diab2==1&male==1],na.rm=T))
[1] 37.27868
```

Example: One-sample t-test for Cholesterol Levels in Diabetic Males

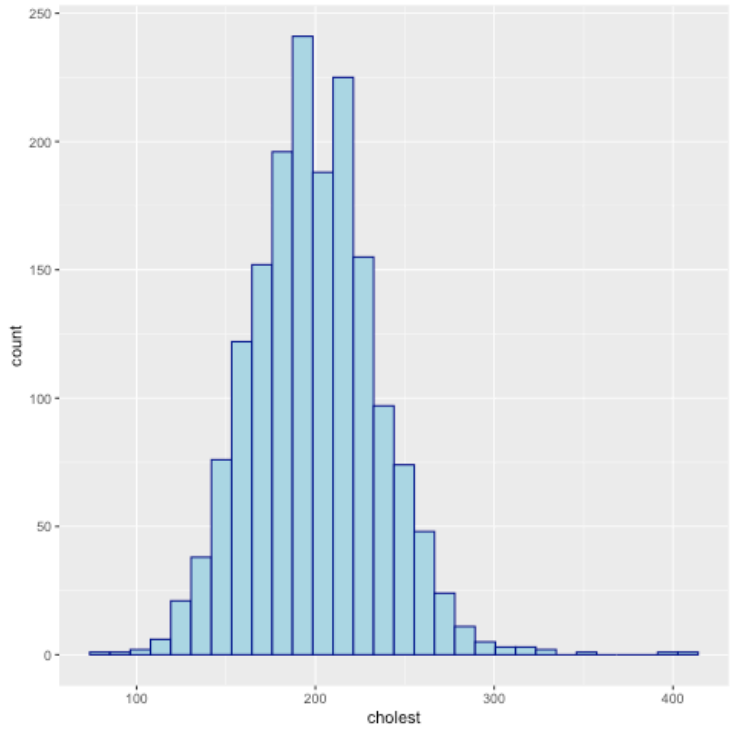
```
> cholestdiabM<-with(inflamm, cholest[diab2==1&male==1])
> summary(cholestdiabM)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
   73.0   168.8   189.0   191.5   213.0   321.0      22
> t.test(cholestdiabM,mu=200,alternative = "two.sided")

        One Sample t-test

data:  cholestdiabM
t = -4.4682, df = 379, p-value = 1.043e-05
alternative hypothesis: true mean is not equal to 200
95 percent confidence interval:
 187.6951 195.2154
sample estimates:
mean of x
 191.4553
```

- The p-value is p=1.04e-05.
- We reject the null hypothesis that mean total cholesterol in elderly diabetic men is 200.
- The observed data are atypical of those that might be obtained when the true mean is 200 (p=1.04e-05).
- We have found significance evidence that elderly diabetic men have mean total cholesterol different from 200 (p=1.04e-05).

Example: One-sample t-test for Cholesterol Levels in Diabetic Males

In our sample of 380 diabetic men, mean cholesterol was 191.46 mg/dl with standard deviation 37.3 mg/dl. With 95% confidence, elderly diabetic men have mean cholesterol between 187.7 mg/dl and 195.2 mg/dl. Our data provide strong evidence that elderly diabetic men have a lower mean cholesterol than 200 mg/dl (p=1.04e-05).

Example: Cholesterol Levels in Non-Diabetic Males




```
> cholestmalediab<-with(inflamm, cholest[diab2==1&male==1])
> summary(cholestnondiabM)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
   78.0   176.0   198.0   199.7   222.0   407.0      20
> sd(cholestnondiabM,na.rm=TRUE)
[1] 35.23686
```

Example: Cholesterol Levels in Non-Diabetic Males

```
> t.test(cholestnondiabM,mu=200,alternative = "two.sided")

        One Sample t-test

data:  cholestnondiabM
t = -0.38199, df = 1693, p-value = 0.7025
alternative hypothesis: true mean is not equal to 200
95 percent confidence interval:
 197.9938 201.3522
sample estimates:
mean of x
  199.673
```

- The p-value is 0.7025.
- We do not reject the null hypothesis that mean cholesterol in non-diabetic men is 200.
- The observed data are not atypical of those that might be obtained when the true mean is 200 (p=0.7025).
- We have not found evidence that non-diabetic elderly men have mean cholesterol different from 200 (p=0.7025).

Example: One-sample t-test for Cholesterol Levels in Non-Diabetic Males

More complete study summary:

In our sample of 1694 elderly men who were not diabetic at the time of enrollment, mean cholesterol was 199.7 mg/dl with standard deviation 35.2 mg/dl. With 95% confidence, non-diabetic elderly men have mean cholesterol between 198.0 mg/dl and 201.3 mg/dl. Our data do not provide evidence that non-diabetic men have mean cholesterol that is different than 200 mg/dl (p=0.70).

P-values and confidence intervals

- There is a strong relationship between p-values and confidence intervals
  - The 95% confidence interval for a parameter θ contains the values of θ for which the null hypothesis $\theta = \theta_0$ would NOT be rejected with a 2-sided alternative.
    - (Assuming same methods are used for the confidence interval and the hypothesis test.)

Confidence Intervals and P-Values for Hypothesis Tests

A level α two-sided test rejects a hypothesis $H_0 : \mu = \mu_0$ exactly when the value of μ_0 falls outside a $(1 - \alpha)$ confidence interval for μ.

For example, consider a two-sided test of the following hypotheses

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

at the significance level $\alpha = .05$.

- If μ_0 is a value inside the 95% confidence interval for μ, then this test will have a p-value greater than .05, and therefore will not reject H_0.
- If μ_0 is a value outside the 95% confidence interval for μ, then this test will have a p-value smaller than .05, and therefore will reject H_0.

More On Constructing Hypothesis Tests

Hypothesis always refer to some population or model, not to a particular outcome. As a result, H_0 and H_a must be expressed in terms of some population parameter or parameters.

H_a typically expresses the effect that we hope to find evidence for. So H_a is usually carefully thought out first. We then set up H_0 to be the case when the hope-for effect is not present.

It is not always clear whether H_a should be one-sided or two-sided, i.e., does the parameter differ from its null hypothesis value in a specified direction.

Note: You are not allowed to look at the data first and then frame H_a to fit what that data show.

Potential Abuses of Tests

In many applications, a researcher constructs a null hypotheses with the intent of discrediting it.

For example:

- H_0: new drug has the same effect as placebo
- H_0: men and women are paid equally

A small p value can help a drug company can get a drug approved by the FDA. Similarly, a researcher may have an easier time publishing his results if the p-value is smaller than 0.05.

Because of that we have to be aware of the following potential abuses:

- Using one-sided tests to make the p-value one-half as big
- Conducting repeated sampling and testing and reporting only the lowest p-value
- Testing many hypothesis or testing the same hypothesis on many different subgroups.

In the last two, even if there is actually no effect, you will probably get at least one small p-value.

Back to P-Values

- The ***p*-value** (p) is the probability of obtaining a result as extreme or more extreme than the observed sample if H_0 is true.
- A good mnemonic: p-value ≈ P(Data | H_0)
- It is very tempting, and very wrong, to interpret a p-value as P(H_0 | Data)

P-Value Interpretation Quiz

A randomized controlled trial of a new treatment led to the conclusion that it is significantly better than placebo with a p-value < 0.05. Which of the following statements is the best description of the results?

(a) The study proved that treatment is better than placebo.

(b) There is less than a 5% chance that the null hypothesis (treatment is no better than placebo) is true.

(c) If the treatment is not effective, there is less than a 5% chance of obtaining the results that were observed in the study.

(d) There is at least a 95% chance the treatment is effective.

P-Values cont.

- p-value ≈ P(Data | H_0)
- A small p-value means there is evidence against the null hypothesis.
- A large p-value means the data do not provide evidence against the null hypothesis. This does NOT mean the data provide evidence that the null hypothesis is true.

P-Values cont.

- Beware over-reliance on p-values!
  - Point estimates and confidence intervals are more useful
- A p-value is a measure of statistical significance
  - P-values do NOT measure whether a result is important, or large, or has clinical or public health implications