BWH - Biostatistics

Introductory Biostatistics for Medical Researchers

Robert Goldman
Professor of Statistics
Simmons University

Issues in Statistical Inference

Monday, November 19, 2018

1. The Chi-Square Test for Independence

The Chi-Square Test for Independence

Breast-Feeding Status and Race

Rows: BreastFed Columns: Race

	Black	Hispanic	White	All
No	17	5	10	32
	68%	42%	32%	47%
Yes	8	7	21	36
	32%	**58%**	**68%**	**53%**
All	25	12	31	68
	100%	100%	100%	100%

The three sample percentages differ. However, might the corresponding population proportions (p_B, p_H, and p_W) be equal and our sample differences simple due to chance?

Research Question

For low-income mothers in Boston, does the likelihood of breast-feeding vary by race?

The Null Hypothesis (H_0)

1. Over all low-income mothers in Boston, the variable—breast-feeding decision—is independent of race.

2. The proportion of low-income mothers in Boston who breast-feed their infant does not vary by race.

3. H_0: $p_B = p_H = p_W$

The Alternative Hypothesis (H_A)

1. Over all low-income mothers in Boston, the variable—breast-feeding decision—is dependent on race.

2. The proportion of low-income mothers in Boston who breast-feed their infant *does* vary by race.

3. H_A: p_B , p_H and p_W are not all equal.

The basic plan in a Chi-Square test is to compare the observed counts we actually obtained (17, 5, 10, 8, 7, 21) with those that we would have **expected** to obtain had the two variables been independent.

Observed

	Black	Hispanic	White	All
No	17	5	10	32
				47%
Yes	8	7	21	36
				53%
All	25	12	31	68

Expected

	Black	Hispanic	White	All
No	11.76	5.65	14.59	32
				47%
Yes	13.24	6.35	16.41	36
				53%
All	25	12	31	68
				100%

The test statistic is $X^2 = \sum \frac{(O - E)^2}{E}$

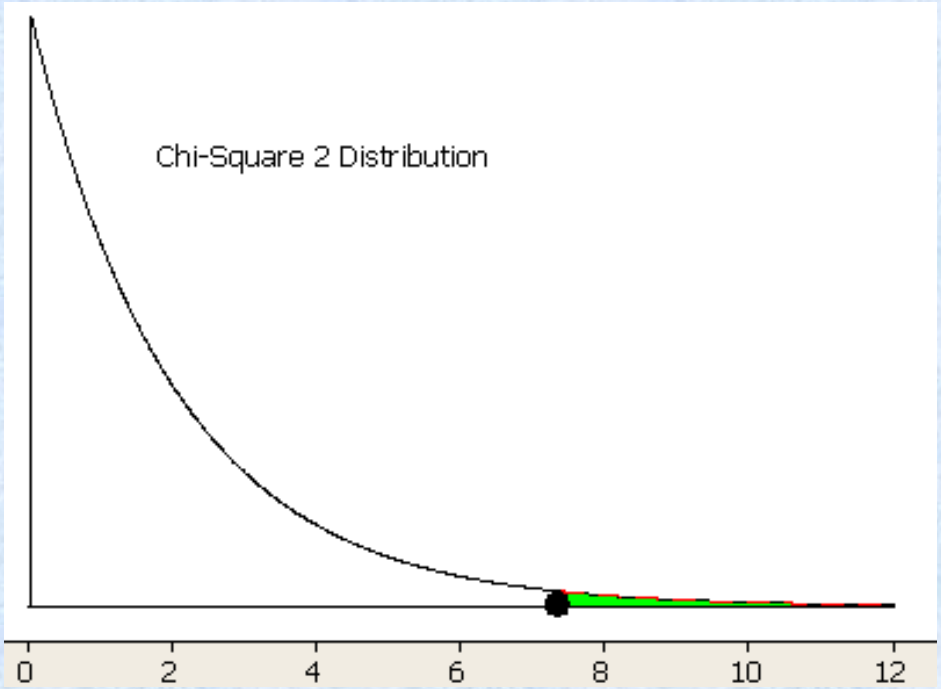
$$= \frac{(17-11.76)^2}{11.76} + \frac{(5-5.65)^2}{5.65} + \frac{(10-14.59)^2}{14.59}$$

$$+ \frac{(8-13.24)^2}{13.24} + \frac{(7-6.35)^2}{6.35} + \frac{(21-16.41)^2}{16.41}$$

$$= 2.3297 + 0.0741 + 1.4431$$

$$+ 2.0708 + 0.0659 + 1.2827 = 7.278$$

Could the differences between the observed and expected counts be attributed to chance?



p-value = shaded area to the right of 7.278 = 0.026.

```
1 - pchisq(7.278, 2)
[1] 0.02627861
```

This tiny p-value tells us that, if H_0: $p_B = p_H = p_W$, is true, a value for X^2 of 7.278 or greater would occur in only 26 of every 1,000 replications of this study.

Rather than concluding that a relatively unlikely outcome occurred when the null is true, when the p-value is small, we prefer to reject the null hypothesis and conclude that the large discrepancy between the observed and the expected counts occurred because the alternative hypothesis is true and that p_B, p_H, and p_W are significantly different.

The data suggest that the percentage of low-income mothers who breast-feed does differ significantly by race.

For an R – by – C contingency table, the number of degrees of freedom for a Chi-Square test is

$$df = (R - 1)(C - 1)$$

For our example, R = 2 and C = 3 and the appropriate number of df is $(R - 1)(C - 1) = (1)(2) = 2$

Once you have obtained any 2 expected values the others can be found by subtraction.

	Black	Hispanic	White	
No				32
Yes				36
	25	12	31	68

```
tally(breastfed ~ race, data = infants)
          race
breastfed Black Hispanic White
      No     17        5    10
      Yes     8        7    21
```

```
round(tally(breastfed ~ race, data = infants,
 format = "percent"), 1)
          race
breastfed Black Hispanic White
      No   68.0     41.7  32.3
      Yes  32.0     58.3  67.7
```

```
c <- chisq.test(infants$breastfed,
   infants$race)
c
     Pearson's Chi-squared test

data:  infants$breastfed and infants$race
X-squared = 7.2664, df = 2, p-value = 0.02643
```

```
round(c$expected, 1)
                  infants$race
infants$breastfed Black Hispanic White
              No   11.8      5.6  14.6
              Yes  13.2      6.4  16.4
```

2. General Features of Hypothesis Testing

General Features of Hypothesis Testing

1. There are always two hypotheses, the null hypothesis and the alternative hypothesis.

- The null invariably expresses the absence of an effect.

- The researcher almost always favors the alternative

2. In some contexts the researcher may select either a one-sided or a two-sided alternative hypothesis.

- For t-tests one can use either a one-sided or a two-sided alternative.

- Some tests such as Chi-Square tests (and F-tests in ANOVA) are intrinsically two-sided.

3. Hypotheses must be formulated *before* seeing the data.

4. Hypotheses (and conclusions about them) are always about features of the **population(s)** (μ_1, μ_2, p_W, p_H, p_B, etc). We use the sample data to test the credibility of the null hypothesis.

5. Although we have two hypotheses, we test only one of the two, H_0.

6. We don't treat the two hypotheses equally. In fact, we begin the process of testing the null hypothesis by assuming that it is true.

We only reject the null hypothesis (in favor of H_A) if the data provides compelling evidence that the null is not true.

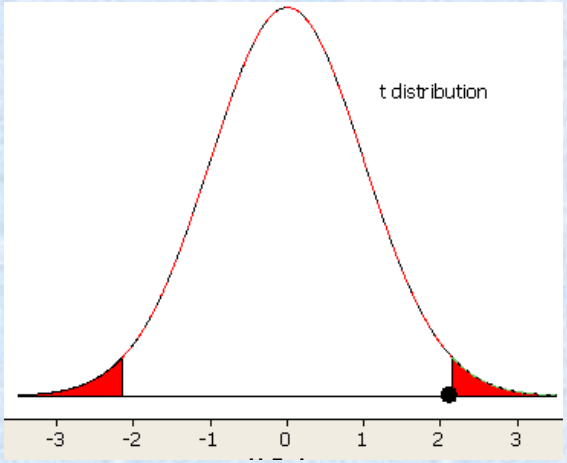
7. We compute a test statistic, the magnitude of which, reflects how far the sample results differ from those that would be expected if the null hypothesis were true.

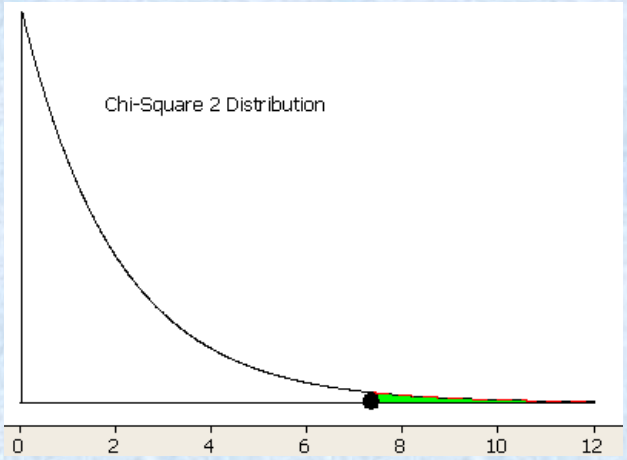
$$t_0 = \frac{(\bar{X}_1 - \bar{X}_2) - 0}{SE(\bar{X}_1 - \bar{X}_2)} = \frac{(\bar{X}_1 - \bar{X}_2) - 0}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

$$X^2 = \sum \frac{(O - E)^2}{E}$$

8. A test statistics is only useful if we know the **sampling distribution** (the pattern of behavior of the test statistic in repeated samples) of that statistic when the null hypothesis is true. Areas under these sampling distributions can be interpreted as probabilities.

9. The area(s) under the appropriate distribution beyond the observed value of the test statistic is called the **p-value**.





10. The p-value is a measure of the strength of the evidence against the null hypothesis. The smaller the p-value the stronger the evidence against the null hypothesis and the more inclined we should be to reject H_0 in favor of H_A.

11. The p-value is **the probability of getting our sample result (or a result even more extreme), assuming H_0 is true**.

The p-value is <u>not</u> the P(Null hypothesis is true)

12. Interpreting the size of the p-value

p-value

Convincing Considerable Suggestive Some None

Evidence against the null

13. The Level of Significance

From our t test from last week:

$H_0: \mu_1 - \mu_2 = 0$ $\quad$ $H_A: \mu_1 - \mu_2 \neq 0$

$\bar{X}_1 - \bar{X}_2 = 7.96$ $\quad$ $t = 2.14$ $\quad$ p-value = 0.038

$\bar{X}_1 - \bar{X}_2 = 3.96$ $\quad$ $t = 1.066$ $\quad$ p-value = 0.292

When the difference in sample means was 7.96 ounces, the p-value (for the two-sided alternative) was 0.038 and we rejected the null hypothesis.

When the difference in sample means was 3.96 ounces, the p-value (for the two-sided alternative) was 0.292 and we did not reject the null hypothesis.

What if the p-value had been 0.15; or 0.09; or 0.04?

How small does the p-value have to be before we are willing to reject the null hypothesis?

How unusual does our sample result have to be, assuming the null hypothesis is true, before we reject the null hypothesis?

We call the value at which we start to reject the null hypothesis the **level of significance** and denote it by the symbol **α**. Typically, $\alpha = 0.05$ or less frequently, 0.01.

If the p-value = 0.038 we will reject the null hypothesis at the 5% level of significance.

If the p-value = 0.38 we will not reject the null hypothesis at the 5% level of significance.

Classroom Exercise

At what point do you begin to doubt that the coin is fair?

Number of Heads	Probability	Percent of Students
1	0.5	0
2	0.25	1%
3	0.125	6%
4	0.0625	28%
5	0.03125	39%
6	0.015625	14%
7	0.0078125	10%
8	0.00390625	2%

Weighted Average of the probabilities = 0.043

14. Errors in Hypothesis Testing and Power Analysis

In a recent randomized, double-blinded, placebo-controlled, trial in Minnesota, researchers were interested in evaluating the effectiveness of an influenza vaccine at reducing the incidence of upper respiratory disease.….

The researchers were interested in testing:

H_0: $p_p - p_v = 0$ against $p_p - p_v \neq 0$ where

p_p = P(URI given placebo)

p_v = P(URI given vaccine)

The **true effect size** is then $E = p_p - p_v$

If you insist on making a decision based on sample data, you must recognize that your decision may be incorrect. Whether or not your decision is in error depends on the 'truth' in the population.

		Truth	
		H_0: is true	Some H_A is true
		$p_p - p_v = 0$	$p_p - p_v \neq 0$
Decision based on data	Reject H_0 because p value $< \alpha$	*Type I error*	*Fine*
	Do not reject H_0 because p value $> \alpha$	*Fine*	*Type II error*

P(Type I error) = P(Rejecting H_0 when H_0 is true)

$= \alpha$

P(Type II error) = P(Not rejecting H_0 when H_A is true)

$= \beta$

$1 - \beta$ = P(Rejecting H_0 when H_A is true)

= **Power** of the test

The researcher selects the value for α but the values for β (and $1 - \beta$) depend on the value for α, the sample size(s), and, most crucially, on the **true effect size** (the true value for $p_p - p_v$).

A **power analysis** involves examining the tradeoffs between $1 - \beta$, the effect size, and the sample size.

In general, the researcher selects the value for α but the values for β (and $1 - \beta$) depend on the value for α, the sample size(s), and, most crucially, on the **effect size** (the true value for $p_p - p_v$).

The researchers performed a power analysis, exploring the relationship between sample size, the power of the test and the *desired* effect size ($D = p_1 - p_2$). They assumed $\alpha = 0.05$ and a two-sided test. As a result, they decided that they needed a total of n = 400 in each group.

The software, R makes it very easy to make power calculations in the context of comparing two proportions:

The user must specify:

(a) guesses for p_p and p_v

(b) either the desired sample size (in which case the output will give you power) or the desired power (in which case the output will give you the needed sample size).

By default, R assumes a two-sided test and $\alpha = 0.05$.

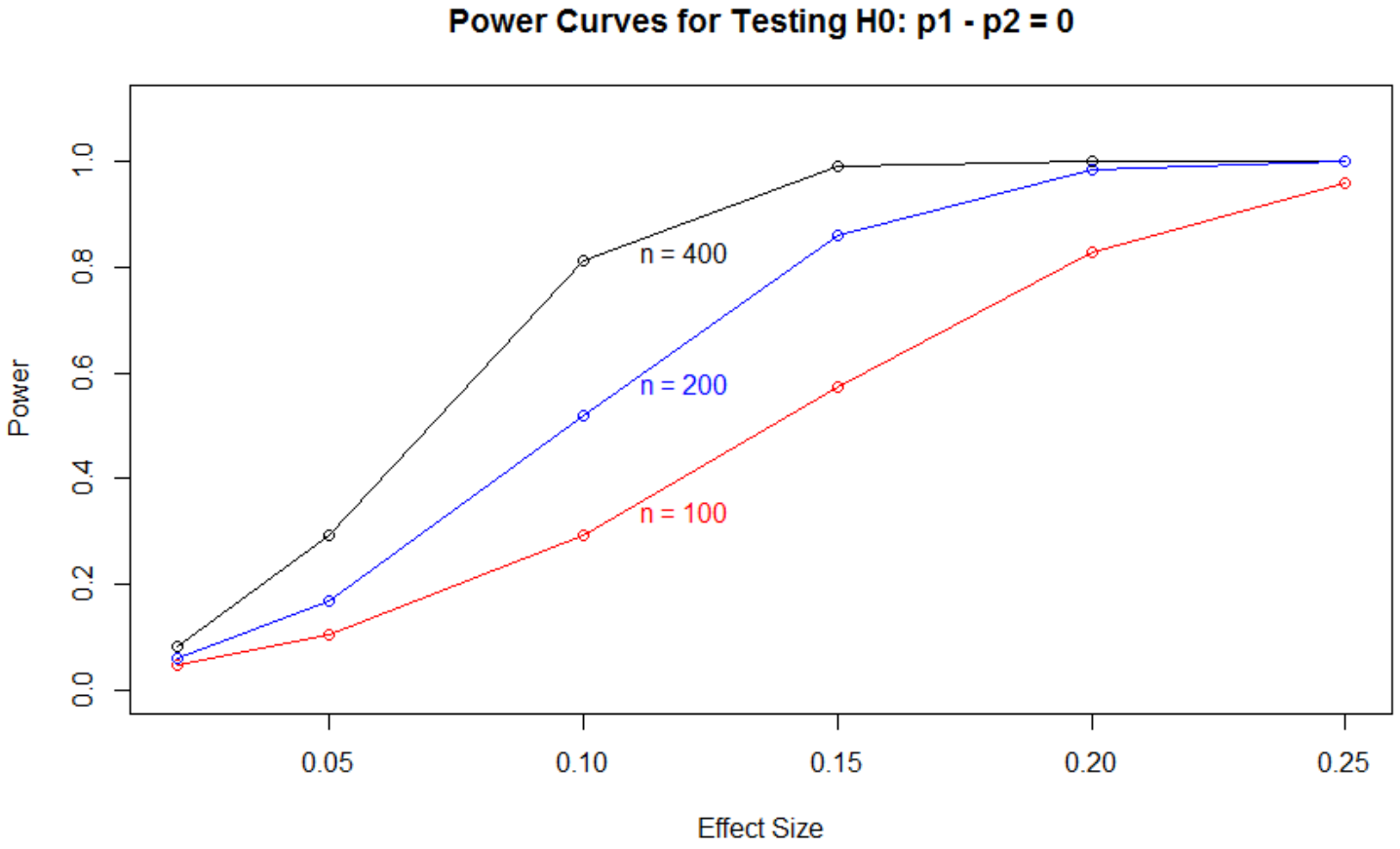
```
power.prop.test(n = NULL, power = 0.8,
   p1 = 0.5, p2 = 0.4)

     Two-sample comparison of proportions power
calculation

              n = 387.3385
             p1 = 0.5
             p2 = 0.4
      sig.level = 0.05
          power = 0.8
    alternative = two.sided

NOTE: n is number in *each* group



power.prop.test(n = 200, power = NULL,
   p1 = 0.5, p2 = 0.4)

     Two-sample comparison of proportions power
calculation

              n = 200
             p1 = 0.5
             p2 = 0.4
      sig.level = 0.05
          power = 0.5200849
    alternative = two.sided

NOTE: n is number in *each* group
```

Table entries are values for power = $1 - \beta$ $p = 0.3$

Effect Size	Number in each group		
$(p_p - p_v)$	n = 100	n = 200	n = 400
0.02	0.050	0.064	0.091
0.05	0.121	0.200	0.353
0.10	0.371	0.638	0.906
0.15	0.722	0.951	0.999
0.20	0.948	0.999	1
0.25	0.998	1	1

Table entries are values for power = $1 - \beta$ $p = 0.5$

Effect Size			
$(p_p - p_v)$	n = 100	n = 200	n = 400
0.02	0.047	0.059	0.082
0.05	0.105	0.169	0.293
0.10	0.294	0.520	0.813
0.15	0.574	0.861	0.991
0.20	0.828	0.985	1
0.25	0.960	1	1




```
e <- c(0.02, 0.05, 0.1, 0.15, 0.2, 0.25)
Power100 <- c(0.047, 0.105, 0.294, 0.574, 0.828, 0.960)
Power200 <- c(0.059, 0.169, 0.520, 0.861, 0.985, 1)
Power400 <- c(0.082, 0.293, 0.813, 0.991, 1, 1)
plot(Power100 ~ e, type = "o", col = "red", ylim = c(0, 1.1),
  xlab = "Effect Size", ylab = "Power",
  main = "Power Curves for Testing H0: p1 - p2 = 0")
lines(Power400 ~ e, type = "o", col = "black")
lines(Power200 ~ e, type = "o", col = "blue")
text(.12, .34, paste("n = 100"), col = "red")
text(.12, .58, paste("n = 200"), col = "blue")
text(.12, .83, paste("n = 400"), col = "black")
```

When we compare groups, why is it better to have the same number in each group?

Because any configuration other than equal sample sizes provides either less precision (in the case of estimation) or less power (in the case of hypothesis testing).

Suppose the sample size calculation calls for 200 subjects per group/sample.

n_1	n_2	M of E for 95% CI for $p_1 - p_2$	Power if $p_1 - p_2 = .1$
200	200	0.0980	0.5160
180	220	0.0985	0.5120
150	250	0.1012	0.4906
120	280	0.1069	0.4495
100	300	0.1132	0.4095
50	350	0.1482	0.2620

Suppose, again, that the sample size calculation calls for 200 subjects per group/sample, but in one group it is possible to recruit only n_1 subjects ($n_1 < 200$).

How many subjects (n_2) need to be recruited in the other group in order to obtain the same precision/power that we would have achieved with 200 in each?

n_1	n_2	Total	
200	200	400	Not possible
180	225	405	
140	350	490	
120	600	720	
102	5100	5202	
100	∞		

Final Thoughts on Sample Size Determination

- Sample-size determination involves collaboration between the researcher and the statistician.
- Determining the sample size must be done before seeking funding and certainly before recruiting subjects.
- In many cases sample size is determined by resource limitations.
- There may be numerous estimates of sample size - one for each response variable.
- Don't forget that you may have a sample size that is adequate for the entire group but not adequate for sub-groups.
- Sample size calculations should take into account anticipated non-response or attrition.

An Example

In a recent randomized, double-blinded, placebo-controlled trial in Minnesota, 800 healthy volunteers were randomly assigned to receive an influenza vaccine or a placebo. The subjects were recruited in October and November. In the following April the number of episodes of upper respiratory illnesses was recorded for each subject.

Here are the (sample) results:

	Placebo	Vaccine	Difference	p
n	400	400		
$\bar{X}$	**1.40**	**1.05**	**0.35**	**0.0001**
S	1.56	1.45		

15. Good researchers always provide the p-value!

Good Practice	Placebo	Vaccine	p
n	400	400	
Mean number URIs	1.40	1.05	0.001

Bad Practice	Placebo	Vaccine	p
n	400	400	
Mean number URIs	1.40	1.05	$p < 0.05$

Poor Practice	Placebo	Vaccine	p
n	400	400	
Mean number URIs	1.40	1.05	*

* $p < 0.05$

3. Some Important Tests of Independence

Some Important Tests of Independence

Explanatory Variable	**Response Variable**	**Test**
Treatment (Placebo or Vaccine	Number of URI	Two-sample t test
Qualitative	Quantitative	

Explanatory (Independent) Variable

Response (Dependent) Variable	Qualitative	Quantitative
Qualitative	Chi-Square test for Independence	Chi-Square test in Logistic Regression
Quantitative	1. Paired t test 2. Two-sample t test 3. F test in One-Way ANOVA	t test in Linear Regression

Explanatory Variable	Response Variable	Test
Treatment (Placebo, Vaccine1 & Vaccine2)	Number of URI	F test in One-Way ANOVA
Qualitative	Quantitative	

Response (Dependent) Variable	Explanatory (Independent) Variable: Qualitative	Explanatory (Independent) Variable: Quantitative
Qualitative	Chi-Square test for Independence	Chi-Square test in Logistic Regression
Quantitative	1. Paired t test 2. Two-sample t test 3. F test in One-Way ANOVA	t test in Linear Regression

Qualitative Response Variable

Explanatory Variable	Response Variable	Test
Treatment (Placebo or Vaccine)	Whether or not at least one URI	Chi-Square test for Independence
Qualitative	Qualitative	

Response (Dependent) Variable	Explanatory (Independent) Variable: Qualitative	Explanatory (Independent) Variable: Quantitative
Qualitative	Chi-Square test for Independence	Chi-Square test in Logistic Regression
Quantitative	1. Paired t test 2. Two-sample t test 3. F test in One-Way ANOVA	t test in Linear Regression

Explanatory Variable	Response Variable	Test
Dose of Vaccine	Number of URIs	t test in Linear Regression
Quantitative	Quantitative	

Explanatory (Independent) Variable

Response (Dependent) Variable	Qualitative	Quantitative
Qualitative	Chi-Square test for Independence	Chi-Square test in Logistic Regression
Quantitative	1. Paired t test 2. Two-sample t test 3. F test in One-Way ANOVA	t test in Linear Regression

Explanatory Variable	Response Variable	Test
Dose of Vaccine	Whether or not at least one URI	Chi-Square test in Logistic Regression
Quantitative	Qualitative	

Explanatory (Independent) Variable

Response (Dependent) Variable	Qualitative	Quantitative
Qualitative	Chi-Square test for Independence	Chi-Square test in Logistic Regression
Quantitative	1. Paired t test 2. Two-sample t test 3. F test in One-Way ANOVA	t test in Linear Regression

4. Misconceptions and Misuses of Hypothesis Testing

Hypothesis Testing: Misconceptions and Misuses

1. The definition of the p-value

The p-value is the probability of getting our sample result (or a result even more extreme), assuming H_0 is true.

The p-value is not the probability of the null hypothesis being true given our sample result!

2. Not Rejecting the Null

Deciding not to reject the null hypothesis (because of a p-value that is not sufficiently small) does not mean that the null hypothesis is true. It means that we do not have sufficient evidence to reject the null. **Failure to reject the null does not mean that the treatments are equivalent!**

3. Hypothesis Testing and Causality

Rejecting the null hypothesis does not mean that you have established a causal relationship between the explanatory and the response variable.

Statistical Inferences and the Design of Studies

	Placebo	Vaccine	Difference	p-value
n	400	400		
Mean	1.40	1.05	**0.35**	0.0001

How do we account for the difference 0.35?

(a) Treatment effect (that is, $\mu_p - \mu_v >> 0$)

(b) Chance

(c) Confounding variables

Suppose the data above was obtained not in a randomized study but in an observational study in which a group of 400 healthy people who had asked for a flu shot were compared with a group of 400 healthy people who had not.

(a) Treatment effect

(b) Chance

(c) Confounding variables

4. **Statistical Significance vs. Clinical Significance**

If the p-value < α we say that the results are **statistically significant at the α level of significance**. However, statistical significance simply indicates that a result is unlikely to have occurred by chance if the null hypothesis is true.

A statistically significance result is not the same thing as a *clinically* significant result. Similarly, a result that is not statistically significant may be of considerable clinical significance.

Statistical Significance:

Compare p-value with α

Clinical Significance:

Compare the sample result ($\bar{X}_1 - \bar{X}_2 = 7.96$) with the null value ($\mu_1 - \mu_2 = 0$)

1.Treatment	n	Mean	p-value
Placebo	400	1.40	0.001
Vaccine	400	1.05	

Statistical Significance? Clinical Significance?

2. Treatment	n	Mean	p-value
Placebo	20	1.40	0.461
Vaccine	20	1.05	

Statistical Significance? Clinical Significance?

3. Treatment	n	Mean	p-value
Placebo	20	1.40	0. 967
Vaccine	20	1.38	

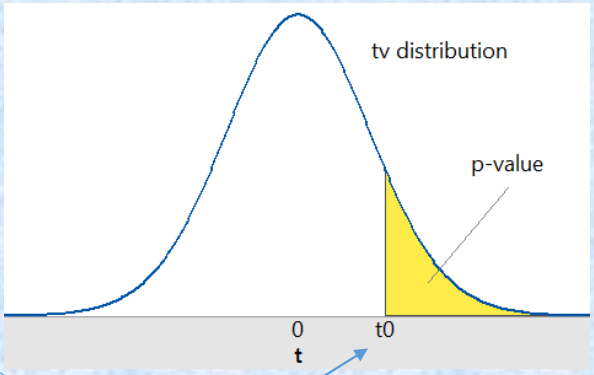
Statistical Significance? Clinical Significance?

4. Treatment	n	Mean	p-value
Placebo	45,000	1.40	0.0466
Vaccine	45,000	1.38	

Statistical Significance? Clinical Significance?

5. What factors affect the p-value?

$$t_0 = \frac{(\bar{X}_1 - \bar{X}_2) - 0}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$



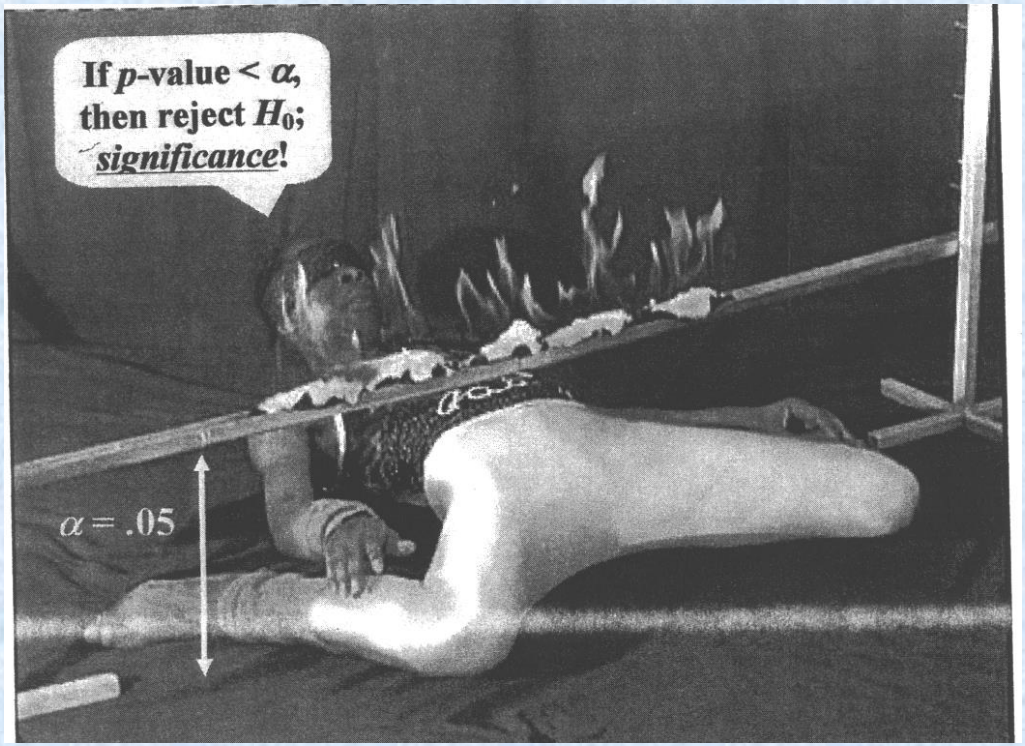
The *larger* the value for t_0 the *smaller* the p-value.

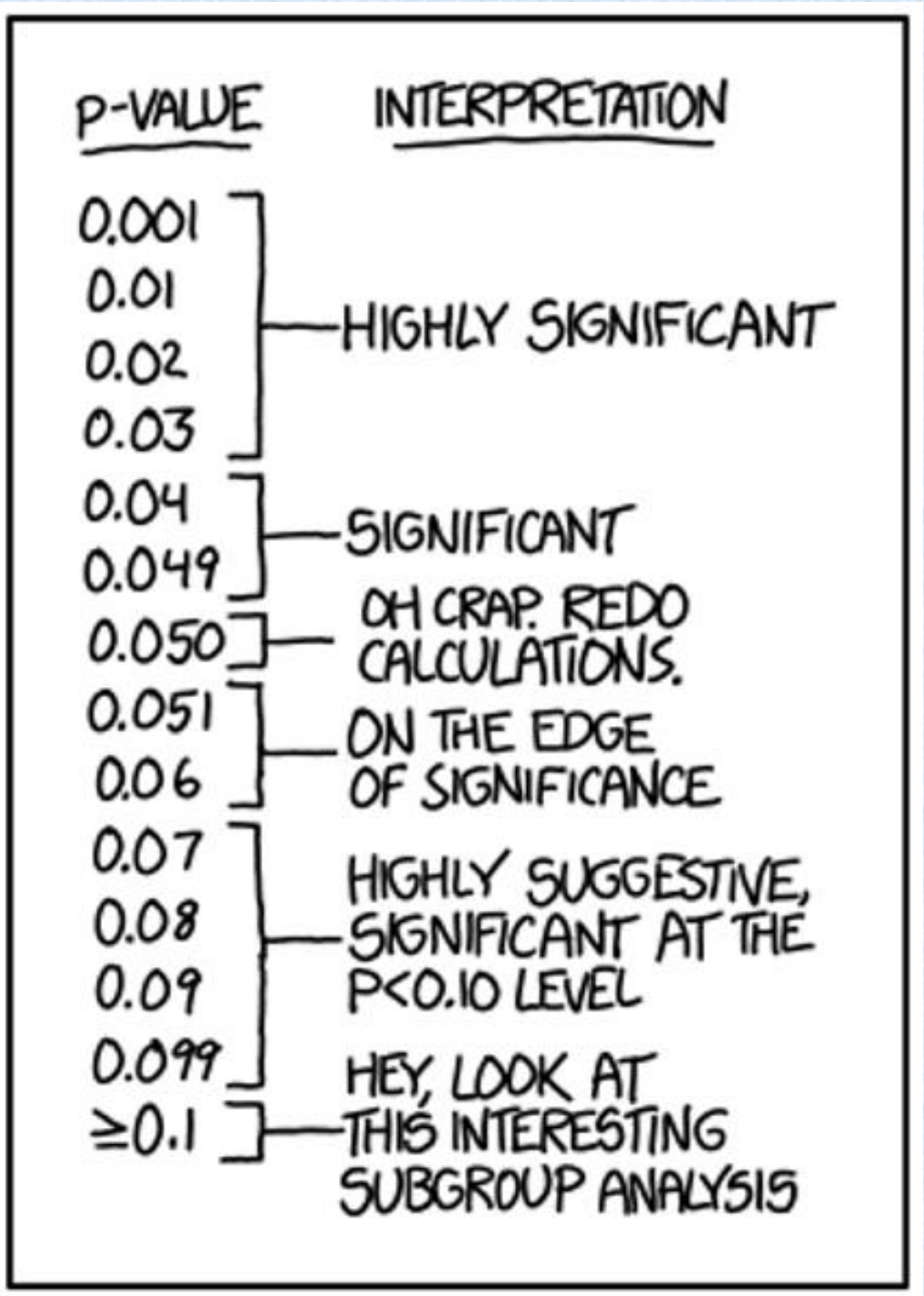
All other factors remaining constant:

- As $\bar{X}_1 - \bar{X}_2$ increases, the value for t_0 increases and the p-value decreases in value
- As S_1 and S_2 increases in value, the value for t_0 decreases and the p-value increases in value
- As n_1 and n_2 increase in value, the value for t_0 increases and the p-value decreases in value

6. Slavish Devotion to α = 0.05

One of the factors that has given hypothesis testing a bad press in recent years is the view that you must always make a decision; more specifically the almost slavish attachment of researchers and medical journals to a level of significance of 5% has led to a distortion in published medical research.




P-VALUE
INTERPRETATION
0.001
0.01
0.02
0.03
HIGHLY SIGNIFICANT
0.04
0.049
SIGNIFICANT
0.050
OH CRAP. REDO CALCULATIONS.
0.051
0.06
ON THE EDGE OF SIGNIFICANCE
0.07
0.08
0.09
HIGHLY SUGGESTIVE, SIGNIFICANT AT THE P<0.10 LEVEL
0.099
≥0.1
HEY, LOOK AT THIS INTERESTING SUBGROUP ANALYSIS

An Extreme Example

1. Suppose that we are interested in comparing a new influenza vaccine that it is hoped will be superior to the current vaccine in reducing the mean number of URIs.

2. Suppose further, that the new vaccine is, in reality, no better than the current vaccine; that is $\mu_{New} = \mu_{Current}$.

3. The vaccines are compared in each of 20 independent experiments at sites throughout the country. A level of significance of $\alpha = 0.05$ is used at each site.

4. In 19 of the 20 experiments the researchers (rightly) fail to reject the null hypothesis. At the 20^{th} site the p-value is 0.047; the researchers at that site write up there results and submit them to the NEJM. The NEJM accepts this article.

5. Citing the article, the developers of the new vaccine apply to the FDA for approval to mass produce the vaccine!

7. Multiple Tests

All of our elaborate hypothesis testing structure is based on collecting data and performing a single test. But typically, we perform many tests on a data set; there are often multiple response/outcome variables, and we frequently compare treatments for many subgroups.

In a single test the probability of falsely rejecting the null hypothesis is α (often 0.05). But what if we perform K tests? What is relevant is the K-test wide probability of at least one false reject of the null.

Suppose, in fact, the null hypothesis is true for all K tests and we use a level of significance of α for each test.

P(at least one false rejection of the null)

= 1 - P(we don't reject the null for all K tests)

= $1 - (1 - \alpha)^K$

If $K = 20$ and $\alpha = 0.05$

P(at least one false rejection of the null)

$= 1 - (1 - \alpha)^K$

$= 1 - (0.95)^{20}$

$= 1 - 0.3585$

$= 0.6415$

Suppose we want to control this overall error rate at 0.05. What should α be?

$0.05 = 1 - (1 - \alpha)^K \cong 1 - (1 - \alpha K) = \alpha K.$

So, $\alpha = 0.05/K$.

If you want to control the overall error rate at 0.05 and K = 20 use $\alpha = 0.05/20 = 0.0025$.

If you want to control the overall error rate at 0.05 and K = 10 use $\alpha = 0.05/10 = 0.005$.

8. Hypothesis Testing and Random Sample(s)

All of traditional statistical inference is predicated on the assumption that the sample(s) were randomly drawn from some well defined population.

In CDC surveys, for example, this assumption is plausible.

But, in most all clinical trials, whether randomized trials or observational studies, the subjects are almost never randomly selected. Still we conveniently forget this random sample requirement when we perform t-tests and Chi-Square tests

Standard practice is to examine the characteristics of the participants and use the results to ‘define’ the population.

5. A Computer-Intensive Approach to Hypothesis Testing

What is wrong with the traditional approach to inference?

- The distributional assumptions underlying the two-sample t test are often violated
- In modern (medical) research subjects are almost never randomly *selected*; once subjects are obtained, the gold standard is random *assignment*!
- The logic of traditional statistical inference is conceptually complex
- Traditional inference in this context forces us to compare means

Two-Sample Permutation Test

H_0: Infant birth-weight is independent of smoking status

H_A: Infant birth-weight is dependent on smoking status. In particular, non-smoking mothers will tend to have heavier infants than smoking mothers.

Infant Birthweight	Smoker	Non-smoker	All
n	19	49	68
ΣX	2056.64	5693.53	7750.2
$\bar{X}$	108.24	116.20	113.97
	$\bar{X}_S$	$\bar{X}_N$	$\bar{X}$

$\bar{X}_N - \bar{X}_S = 7.96$

If the null hypothesis is true, and birth-weight is independent of smoking status then any subset of 19 of the 68 birth-weights can belong to smokers.

The *ideal* behind the permutation test is to compute the difference $\bar{X}_N - \bar{X}_S$ for every possible arrangement/permutation of the 68 birth-weights into 19 'smokers' and 49 'non-smokers' and compute the p-value as the fraction of the differences in means that are greater than or equal to 7.96.

Sadly, there are billions of such permutations and so what we do is compute the sample differences over a large number of randomly selected permutations.

Step 1

Select a random sample of 19 of the 68 infants. These are our ‘smokers’. Now, compute the mean of these 19 infants, $\bar{X}_S$.

Step 2

Repeat step 1 9999 times

1	$\bar{X}_{S1}$		
2	$\bar{X}_{S2}$		
3	$\bar{X}_{S3}$		
4	$\bar{X}_{S4}$		
5	$\bar{X}_{S5}$		
:	:		
:	:		
9999	$\bar{X}_{S9999}$		

Step 3

Now, for each of the 9999 samples of size 19, compute the mean of the other 49 'non-smoking' birth-weights.

$$113.97 = \bar{X} = \frac{19*\bar{X}_S + 49*\bar{X}_N}{68}$$

So,

$$\bar{X}_N = \frac{68*\bar{X} - 19*\bar{X}_S}{49}$$

$$= \frac{7750.2 - 19*\bar{X}_S}{49}$$

	Steps 1, 2	Step 3	
1	$\bar{X}_{S1}$	$\bar{X}_{N1}$	
2	$\bar{X}_{S2}$	$\bar{X}_{N2}$	
3	$\bar{X}_{S3}$	$\bar{X}_{N3}$	
4	$\bar{X}_{S4}$	$\bar{X}_{N4}$	
5	$\bar{X}_{S5}$	$\bar{X}_{N5}$	
:	:	:	
:	:	:	
9999	$\bar{X}_{S9999}$	$\bar{X}_{N9999}$	

Step 4

Compute the 9999 differences $\bar{X}_N - \bar{X}_S$

	Steps 1, 2	Step 3	Step 4
1	$\bar{X}_{S1}$	$\bar{X}_{N1}$	$\bar{X}_{N1} - \bar{X}_{S1}$
2	$\bar{X}_{S2}$	$\bar{X}_{N2}$	$\bar{X}_{N2} - \bar{X}_{S2}$
3	$\bar{X}_{S3}$	$\bar{X}_{N3}$	$\bar{X}_{N3} - \bar{X}_{S3}$
4	$\bar{X}_{S4}$	$\bar{X}_{N4}$	$\bar{X}_{N4} - \bar{X}_{S4}$
5	$\bar{X}_{S5}$	$\bar{X}_{N5}$	$\bar{X}_{N5} - \bar{X}_{S5}$
:	:	:	:
:	:	:	:
9999	$\bar{X}_{S9999}$	$\bar{X}_{N9999}$	$\bar{X}_{N9999} - \bar{X}_{S9999}$

Step 5

Compute the number (k) of differences in means

$(\bar{X}_N - \bar{X}_S)$ that are equal to or greater than or observed difference, 7.96 ounces.

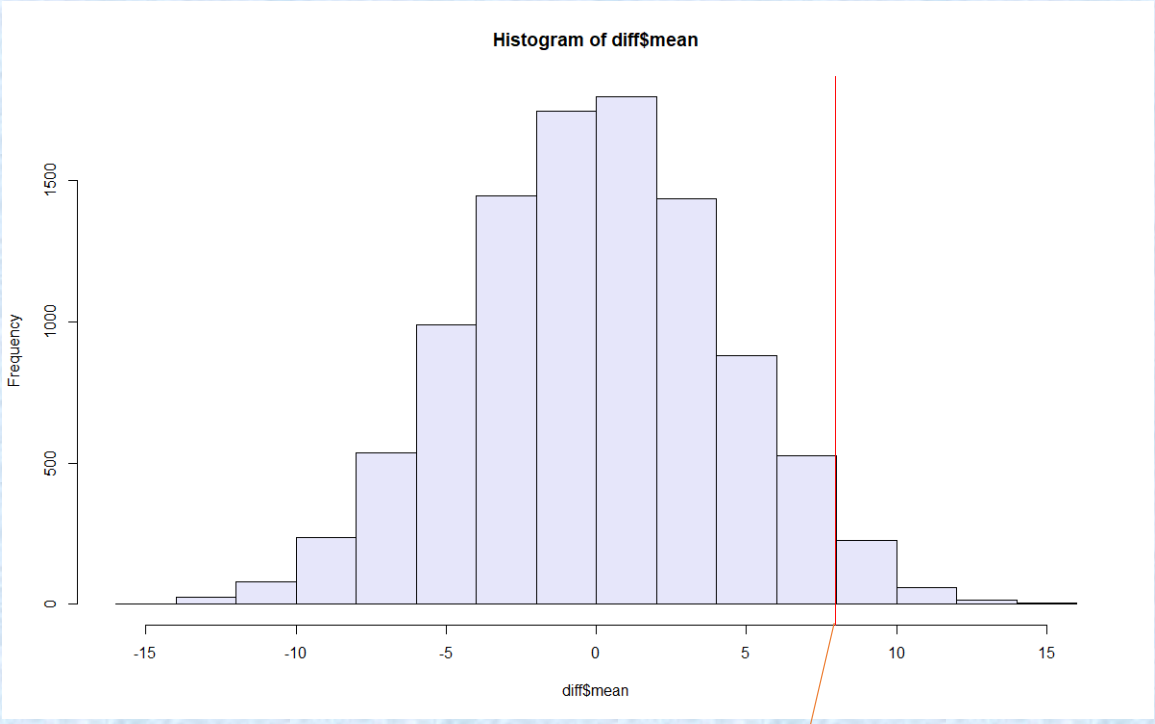
Step 6

We add our actual sample result to the number of samples and as one of the samples with a difference in sample means that is greater than 7.96.

So, compute the p-value as

p-value = (k + 1)/10000

```
xbar_s <- do(9999)*mean(sample(infants$bwt, 19))
xbar_n <- (7746 - 19*xbar_s)/49
diff <- xbar_n - xbar_s
k <- sum(diff >= 7.96)
k
[1] 311
pvalue <- (k + 1)/10000
pvalue
[1] 0.0312
```



7.96

More on Statistical Inference

Questions

1. Several years ago, a multi-center, randomized clinical trial was conducted to test the effectiveness of an influenza vaccine at reducing the incidence of upper respiratory infections (URIs). Approximately 800 healthy adults were randomly assigned to a group that received a flu vaccine or to a group that received a placebo shot. The data are contained in the data frame *u.csv*.

```
> u
    URIs   Group
1      1 Placebo
2      2 Placebo
3      2 Placebo
4      0 Placebo
5      0 Placebo
6      1 Placebo
7      1 Placebo

396    3 Placebo
397    2 Placebo
398    1 Placebo
399    2 Placebo
400    2 Placebo
401    1 Vaccine
402    2 Vaccine
403    1 Vaccine
404    0 Vaccine

796    0 Vaccine
797    0 Vaccine
798    0 Vaccine
799    0 Vaccine
800    0 Vaccine
```

(a) Provide a descriptive summary of these data.

(b) Perform the Shapiro-Wilks test on both samples

(c) We recognize that in large samples, the Normality assumption is unimportant. Perform the two-sided Welch two-sample t test.

(d) Perform the appropriate randomization test and report your conclusion. Use the code below to perform 9999 replications and display the results.

```
pmean <- replicate(9999, mean(sample(u$URIs,
400)))
vmean <- 2*mean(u$URIs) - pmean
diff <- pmean - vmean
pvalue <- (sum(diff >= 0.354)+1)/10000
pvalue
hist(diff, breaks = 30, col = "green")
abline(v = 0.354, col = "red")
```

2. Twenty years ago 40 children were born in the same hospital in Denmark. Researchers classified the children by whether or not they were breast-fed for at least three months. They were interested in whether breast-feeding tends to increase IQ levels. Recently the 40 young adults were given IQ tests. The testers did not know which group a subject belonged to. The results are summarized below as part of the output for a two-sample t test.

```
Two-sample T for IQ

Brstd    N   Mean  StDev  SE Mean
Yes     18  113.4   14.6      3.4
No      22  103.5   15.2      3.2

T-Test of difference = 0 (vs >): T-Value = 2.09
P-Value = 0.0229
```

In the last assignment you concluded that the mean IQ for breast-fed infants was significantly greater than the corresponding mean IQ for non-breast-fed infants. In reaching this conclusion, what type of hypothesis testing error might you have made? State your answer in context.

3. You are planning to perform 25 tests (comparisons) on data that you have collected. You would like the overall probability of at least one Type I error to be 0.1 What (approximate) level of significance should you use for each test?

4. A researcher at Massachusetts General Hospital obtained the systolic blood pressure (SBP) and the sex for 100 low birth-weight infants. She was interested in testing whether or not these two variables are independent.

(a) What is the explanatory variable in this case?

(b) What is the response variable?

(c) What hypothesis test would be the most appropriate starting point in this case?

5. Twenty young women participate in a study of the effect of aspirin on blood clotting time. Ten of the women (randomly selected from the 20) receive a pin prick and the blood clotting time is recorded. A little while later they are given two aspirins. After 30 minutes, the women are given a second pin-prick and the clotting time is measured for the second time. The other ten women are also given two pin-pricks but first starting with aspirin and then without aspirin. We are interested in the increase in clotting time associated with the aspirin.

(a) What is the explanatory variable in this case?

(b) What is the response variable?

(c) What hypothesis test would be the most appropriate starting point in this case?

6. A number of years ago the American Academy of Pediatrics recommended that tetracycline drugs not be used for children under the age of 8. Prior to this recommendation, a two-year study was conducted in Tennessee to investigate the extent to which physicians prescribed this drug in the prior to years. In the study, a random sample of 770 family practice physicians were characterized according to whether the county of their practice was urban, suburban, or rural and by whether they did or did not prescribe tetracycline to at least one child under the age of 8 in the previous year. Are these two variables independent?

(a) What is the explanatory variable in this case?

(b) What is the response variable?

(c) What hypothesis test would be the most appropriate starting point in this case?

7. If you are a dog lover, perhaps having your dog along reduces the effect of stress. To examine the effect of pets in stressful situations, researchers recruited 45 women who said they were dog lovers. Fifteen of the subjects were randomly assigned to each of three groups to do a stressful task (i) alone, (ii) with a good friend present, or (iii) with their dog present. The subject's mean heart rate during the task is one measure of the effect of stress.

(a) What is the explanatory variable in this case?

(b) What is the response variable?

(c) What hypothesis test would be the most appropriate starting point in this case?

The following two questions are multiple choice.

8. When performing a test of significance for a null hypothesis, H_0 against an alternative hypothesis H_A, the p-value is:

a. The probability that H_A is true given the sample data

b. The probability of observing a sample result at least as extreme as that observed if H_0 is true

c. The probability of observing a sample result at least as extreme as that observed if H_A is true

d. The probability that H_0 is true given the sample data

9. A recent editorial in the New York Times reported on a clinical trial in which two different drugs for treating breast cancer in younger women were compared. The editorial contained the phrase "The difference fell just shy of statistical significance, so it remains possible that it occurred by chance, …" Which of the following possible p-values is the most consistent with this phrase?

a. p-value = 0.64

b. p-value = 0.46

c. p-value = 0.064

d. p-value = 0.046