Lecture 7: Comparisons of two groups

Supplementary Reading: Pagano/Gauvreau; Chapter 11

Paired Samples

Suppose we are a conducting a study where we collect two measurements for each subject (e.g. "before" and "after").

A study where two (or more) sets of measurements are collected, *at different times*, for each subject, is called a **longitudinal study**. A study where subjects are measured only at one time point is called a **cross-sectional** study.

The following table shows systolic blood-pressure levels (mm Hg) in 10 women while not using oral contraceptives (baseline) and while using (followup) oral contraceptives.

i	SBP while *not* using OCs	SBP while using OCs	Difference
1	115	128	13
2	112	115	3
3	107	106	-1
4	119	128	9
5	115	122	7
6	138	145	7
7	126	132	6
8	105	109	4
9	104	102	-2
10	115	117	2
Mean	115.6	120.4	4.8
SD	10.31	13.23	4.57

How can we measure the difference in SBP between using and not using oral contraceptives?

Null Hypothesis?

Alternative Hypothesis?

Confidence Interval?

Conclusion?

- Dependent vs. Independent Samples

  Dependent Sample (Paired) - observations that are matched in some way (e.g. pre - and post - test measurements on the same subjects, IQ values in husband-wife pairs, or matched studies)

  Independent Sample - observations from different, non-related groups (e.g. birth-weights of unrelated boys and girls, serum iron levels from a sample of healthy children vs. sick children)

- Dependent samples (paired tests)

  – $\delta = \mu_1 - \mu_2$, δ is the population difference in means

  – $H_0 : \delta = 0$ vs. $H_a : \delta \neq 0$.

	Tests for Dependent Data	
	Known Variance	*Unknown Variance*
Test	Paired normal test	Paired t-test
Test Statistic	$Z = \frac{\bar{d} - \delta}{\sigma_d/\sqrt{n}}$	$\boxed{T = \frac{\bar{d} - \delta}{s_d/\sqrt{n}}}$
Distribution of Test Statistic	standard normal	t distribution with n-1 degrees of freedom

t_{n-1}

  – Similar to a one-sample test
    * Only test one random variable in the null hypothesis
    * Instead of analysis on the two recorded values, we focus on the difference, where the 'population' parameter is the true difference δ

Example: Cereal and LDL

A crossover study was conducted to investigate whether oat bran cereal helps to lower serum cholesterol levels in hypercholesterolemic males. Fourteen such individuals were randomly placed on a diet that included either oat bran or corn flakes. After two weeks their low-density lipoprotein (LDL) cholesterol levels were recorded. Each man was then switched to the alternative diet, and after a second two-week period the LDL cholesterol level of each individual was again recorded. The mean of the differences (corn flake LDL - oat bran LDL) is 0.363 mmol/l with standard deviation of the differences equal to 0.406 mmol/l. Test whether LDL levels are different between the two different diet groups at a 0.05 level.

$\alpha = 0{,}05$

1. What type of test should be performed?

   Two sided, paired t-test because:

   - We are asked to test whether levels are different.
   - Observations are paired, one measurement while eating one cereal, another while eating the other cereal, both on the same person.
   - Population standard deviation is unknown. (σ_d is unknown)

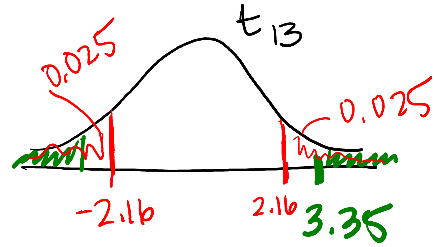
2. Perform the test by hand at a 0.05 level:

   (a) State the null and alternative hypotheses. $H_0: \delta = 0$ $H_A: \delta \neq 0$

   (b) State alpha. $\alpha = 0.05$

   (c) What is the value of the test statistic?

   $$T = \frac{\bar{d} - \delta}{s_d/\sqrt{n}} = \frac{0.363 - 0}{0.406/\sqrt{14}} = 3.35$$



   (d) What is the distribution of the test statistic?

   $$t_{n-1} = t_{14-1} = t_{13}$$

   (e) What is the p-value for the test?

   $$\text{p-value} = 2 \cdot P(t \geq T) = 2 \cdot P(t \geq 3.35) < 2 \cdot 0.005 = 0.01 < \alpha$$

(f) Do you reject or fail to reject the null hypothesis?

p-value < α $\rightarrow$ reject H_0

(g) What conclusion can you draw from this test?

There is evidence to suggest there is a significant difference between LDL chol. after 2 weeks of eating cornflakes vs. 2 weeks of eating oat bran.

(h) How could the confidence interval for the population difference in means be used to test the null hypothesis?

The 95% CI is $(0.13, 0.60)$ (try this on your own as practice). Because the 95% confidence interval does not cover 0 (the population difference in our null hypothesis), we would reject at the $\alpha = .05$ level that $\delta = 0$.

(i) Perform the test using the following R function, and compare your p-value to your answers from above.

```
# Paired t-test
# d: the sample mean of the differences
# s: the sample standard deviations
# n: the same size
# del: the null value for the mean of the differences to be tested for. Default is 0.
# equal.variance: whether or not to assume equal variance. Default is FALSE.
t.pair <- function(d, s, n, del=0)
{
  t <- (d - del)/(s / sqrt(n))
  df <- n-1
  dat <- c(d - del, s, t, 2*pt(-abs(t),df))
  names(dat) <- c("Difference of means", "Std Error", "t", "p-value")
  return(dat)
}

> t.pair(0.363, 0.406, 14, del=0)
Difference of means           Std Error                   t             p-value
        0.363000000         0.406000000         3.345373476         0.005267359
```

- Independent samples (two-sample tests) [Zero is the most common usage]
  - $H_0 : \mu_1 = \mu_2$ or $H_0 : \mu_1 - \mu_2 = 0$
  - $H_a : \mu_1 \neq \mu_2$ or $H_a : \mu_1 - \mu_2 \neq 0$

  Case I: Equal variances

Tests for Independent Data with Equal Variance		
	Known Variance	*Unknown Variance*
Test	Two-sample normal test with equal variances	Two-sample t-test with equal variances
Test Statistic	$Z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\sigma^2[(1/n_1) + (1/n_2)]}}$	$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{s_p^2[(1/n_1) + (1/n_2)]}}$
Distribution of Test Statistic	standard normal	t distribution with $n_1 + n_2 - 2$ degrees of freedom

  - Equation for pooled variance: $s_p^2 = \dfrac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$

  Case II: Unequal Variances

Tests for Independent Data with Unequal Variance		
	Known Variance	*Unknown Variance*
Test	Two-sample normal test with unequal variances	Two-sample t-test with unequal variances
Test Statistic	$Z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{(\sigma_1^2/n_1) + (\sigma_2^2/n_2)}}$	$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{(s_1^2/n_1) + (s_2^2/n_2)}}$
Distribution of Test Statistic	standard normal	t distribution with ν degrees of freedom

  - Equation for degrees of freedom: $\nu = \dfrac{\left[(s_1^2/n_1) + (s_2^2/n_2)\right]^2}{[(s_1^2/n_1)^2/(n_1 - 1) + (s_2^2/n_2)^2/(n_2 - 1)]}$
  - Tests assuming unequal variances are more conservative and are therefore usually chosen unless you are certain that the variances are equal.
  - **Rule of Thumb**: If the ratio of the two sample standard deviations is between 0.5 and 2.0, use the equal variance t-test. Outside of 0.5 to 2.0, use the unequal variance t-test.

Example: Lead Exposure and Neurological Function

A study performed in El Paso, Texas, looked at the association between lead exposure and developmental features in children. There are different ways to quantify lead exposure. One method consists of defining a control group of children whose blood-levels were $< 40\mu g/100$ mL in both 1972 and 1973, and an exposed group of children who had blood-lead levels $\geqslant 40\mu g/100$ mL. An important outcome variable in the study was the number of finger-wrist taps per 10 seconds in the dominant hand, a measure of neurological function. Summary statistics for the outcome variable are in the following table:

	n	$\bar{x}$	s
Control	63	55.1	10.9
Exposed	32	48.4	8.6

$\frac{10.9}{8.6} < 2 \rightarrow$ equal variances

Test for differences in the finger-wrist tap score between the two groups. What do you conclude?

Do you think this study proves or disproves that high levels of lead lead to loss in neurological function? Why or why not?

$H_0: \mu_1 = \mu_2$

$H_A: \mu_1 \neq \mu_2$

$\alpha = 0.05$

$$T = \frac{55.1 - 48.4}{\sqrt{\underbrace{\frac{(63-1)10.9^2 + (32-1)8.6^2}{63+32-2}}_{S_p^2}\left(\frac{1}{63} + \frac{1}{32}\right)}} = 3.03$$

$t_{63+32-2} = t_{93}$

p-value $< 2(0.005) = 0.01 < \alpha \rightarrow$ reject H_0

There is evidence of a significant difference in mean number of finger-wrist taps per 10 seconds comparing children with blood levels $< 40 \mu g/100$ mL and children with blood levels ≥ 40 μg/mL.

Let's look at the same example, except now the standard deviations have been changed. What kind of test should be used now? Perform the test and state your conclusion.

	n	$\bar{x}$	s
Control	63	55.1	10.9
Exposed	32	48.4	4.6

2-sample t-test with unequal variances

$H_0 : \mu_1 = \mu_2$

$H_A : \mu_1 \neq \mu_2$

$\alpha = 0.05$

$$T = \frac{55.1 - 48.4}{\sqrt{\frac{10.9^2}{63} + \frac{4.6^2}{32}}} = 4.19$$

$$df = \nu = \frac{\left[\frac{10.9^2}{63} + \frac{4.6^2}{32}\right]^2}{\left[\left(\frac{10.9^2}{63}\right)^2 / 62 + \left(\frac{4.6^2}{32}\right)^2 / 31\right]} = 90.77 \rightarrow 90$$

$t_\nu = t_{91}$ p-value $< 0.005 < \alpha \rightarrow$ reject H_0

same conclusion as before

To perform the last two examples in R, use the following function.

For the first example, the command is:
`t.test2(55.1, 48.4, 10.9, 8.6, 63, 32, m0=0, equal.variance=TRUE)`

For the second example, the command is:
`t.test2(55.1, 48.4, 10.9, 4.6, 63, 32, m0=0, equal.variance=FALSE)`

```
# m1, m2: the sample means
# s1, s2: the sample standard deviations
# n1, n2: the same sizes
# m0: the null value for the difference in means to be tested for. Default is 0.
# equal.variance: whether or not to assume equal variance. Default is FALSE.

t.test2 <- function(m1,m2,s1,s2,n1,n2,m0=0,equal.variance=FALSE)
{
  if( equal.variance==FALSE )
  {
    se <- sqrt( (s1^2/n1) + (s2^2/n2) )
    # welch-satterthwaite df
    df <- ( (s1^2/n1 + s2^2/n2)^2 )/( (s1^2/n1)^2/(n1-1) + (s2^2/n2)^2/(n2-1) )
  } else
  {
    # pooled standard deviation, scaled by the sample sizes
    se <- sqrt( (1/n1 + 1/n2) * ((n1-1)*s1^2 + (n2-1)*s2^2)/(n1+n2-2) )
    df <- n1+n2-2
  }
  t <- (m1-m2-m0)/se
  dat <- c(m1-m2, se, t, 2*pt(-abs(t),df))
  names(dat) <- c("Difference of means", "Std Error", "t", "p-value")
  return(dat)
}

> t.test2(55.1, 48.4, 10.9, 8.6, 63, 32, m0=0, equal.variance=TRUE)
Difference of means           Std Error                   t
        6.700000000         2.212283081         3.028545514
            p-value
        0.003180431

> t.test2(55.1, 48.4, 10.9, 4.6, 63, 32, m0=0, equal.variance=FALSE)
Difference of means           Std Error                   t
       6.700000e+00        1.595971e+00        4.198072e+00
            p-value
       6.269657e-05
```

Lecture 8: ANOVA

Supplementary Reading: Pagano/Gauvreau; Chapter 12

Motivation

So far we have studied testing the mean of a single population and then the means of two populations. What happens when we have more than two populations? In medicine we are sometimes faced with investigating situations such as lung cancer where curative treatments are not forthcoming and one has to investigate a large number of potential treatments, and there is some savings in time and effort by doing them simultaneously.

There are several ways to approach this problem, but we will first concentrate on the simple one where we want to test the single hypothesis that the means of each of the populations are equal to each other.

- Data: we start with k independent random samples. For each we have a sample size, a sample mean and a sample standard deviation.

Group	Sample Size	Mean	SD
1	n_1	$\bar{x}_1$	s_1
2	n_2	$\bar{x}_2$	s_2
3	n_3	$\bar{x}_3$	s_3
4	n_4	$\bar{x}_4$	s_4
5	n_5	$\bar{x}_5$	s_5
$\vdots$	$\vdots$	$\vdots$	$\vdots$
k	n_k	$\bar{x}_k$	s_k

  Note: Each of the samples also have corresponding "true" population parameters - μ_k and σ_k

- Analysis of variance, or ANOVA, is an extension of the two-sample equal variance t-test to $k > 2$ groups.

$$H_0 : \mu_1 = \mu_2 = \ldots = \mu_k$$

H_A : at least one pair of population means differ.

- The basic idea: compare two kinds of variation.

  (a) "Within-group variation" - variation of the individual values around the group mean. A weighted average of each of the sample variances.

  $$s_W^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2 + \ldots + (n_k-1)s_k^2}{n_1+n_2+\ldots+n_k-k}$$

  (b) "Between-group variation" - variation of the group means around the overall mean

  $$s_B^2 = \frac{(\bar{x}_1-\bar{x})^2 n_1 + \ldots + (\bar{x}_k-\bar{x})^2 n_k}{k-1}$$

  Where

  $$\bar{x} = \frac{n_1\bar{x}_1 + \ldots + n_k\bar{x}_k}{n_2+n_2+\ldots+n_k}$$

- Assumptions of ANOVA

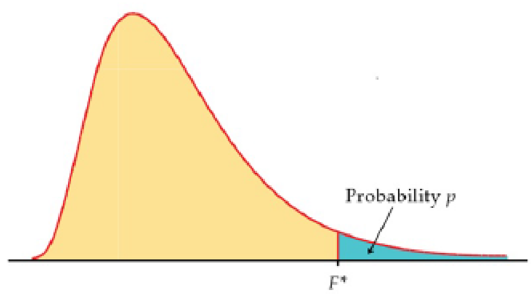
  1. Samples from the k populations are independent.
  2. Samples from the k populations are normally distributed or sample size is large.
  3. Standard deviations in the k populations are equal. i.e., $\sigma_1 = \sigma_2 = \ldots = \sigma_k$.
     – Rule of Thumb- if the largest standard deviation is not more than twice the smallest, then okay to proceed with ANOVA.

- Test Statistic

$$F_{k-1,n-k} = \frac{s_B^2}{s_W^2}$$

- Two types of degrees of freedom.
  - numerator: $k-1$ (corresponds to the df for variation **between** groups) where k is the number of groups.
  - denominator: $n-k$ (corresponds to the df for variation **within** groups) where n is the total number of observations

- The F-statistic **cannot** assume negative values (do not double the p-value)



- Bonferroni Correction

  Our original overall hypothesis was that all the means are equal. So, any departure from this overall equality could be the cause of us rejecting the whole (rejecting H_0). It would be interesting to find out the cause of rejecting H_0. One way to do this, is to perform all pairwise hypothesis tests comparing the means of each pair of groups. However, doing multiple tests requires us to preserve our Type I error, and a correction must be made.

  Our new Type I error becomes:

$$\alpha^* = \frac{\alpha}{\binom{k}{2}}$$

  where $\binom{k}{2}$ denotes the total number of possible pairwise comparisons.

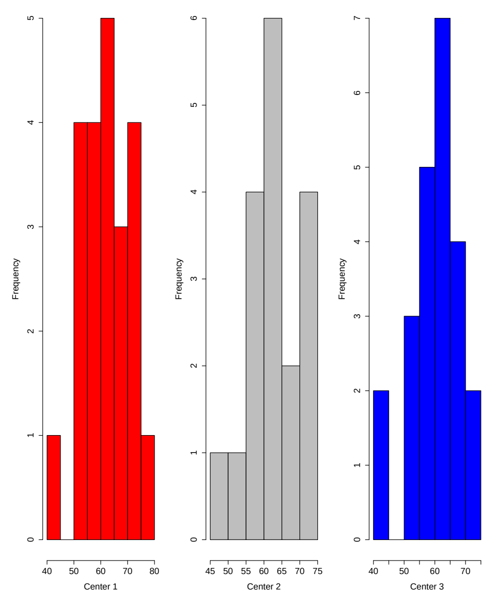
Example: We are interested in examining data from a study (discussed in the text) that investigates the effect of carbon monoxide exposure on patients with coronary artery disease, where baseline measurements of pulmonary function were examined across medical centers. Another characteristic that you might wish to investigate is age. We have the following data:

$K=3$

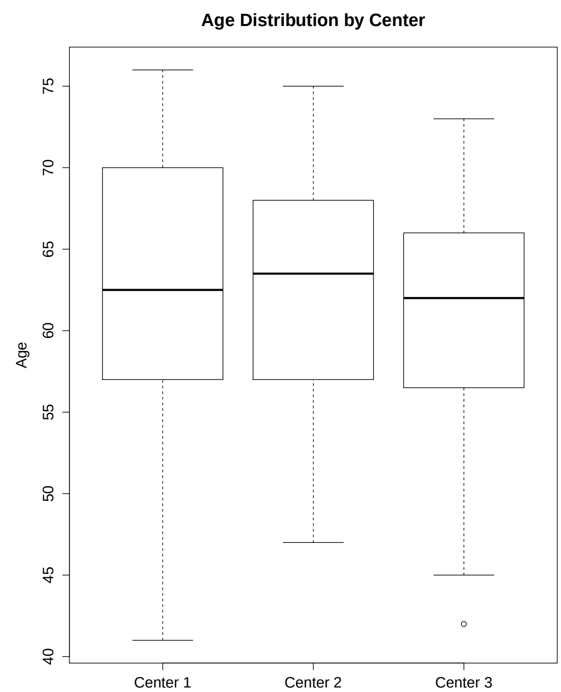
Medical Center	Summary of age (years) Mean	Std. Dev.	Freq.
1	62.55 $\bar{x}_1$	8.67 s_1	22 n_1
2	63.28 $\bar{x}_2$	7.79 s_2	18 n_2
3	60.83 $\bar{x}_3$	8.00 s_3	23 n_3
Total	62.13 $\bar{x}$	8.12	63 n

(a) Let's examine the histogram and boxplots of age for each center. Why is ANOVA an appropriate method for analyzing this data?

```
# histograms of age by center
par(mfrow = c(1,3))
hist(data$age[1:22], xlab = "Center 1", main = "", col = "red")
hist(data$age[23:40], xlab = "Center 2", main = "", col = "gray")
hist(data$age[41:63], xlab = "Center 3", main = "", col = "blue")
```



```
# boxplots of age by center
boxplot(data$age~data$center, names = c("Center 1", "Center 2", "Center 3"),
ylab="Age", main="Age Distribution by Center")
```



From the boxplots and histograms, it seems reasonable to assume that the data are approximately normally distributed. Using the rule of thumb, we see that the greatest standard deviation, 8.67 is not more than twice the smallest standard deviation (7.79) so it is ok to assume our population standard deviations are equal. Finally, the centers represent 3 independent groups.

(b) Let's analyze the data. What are the null and alternative hypotheses? What is alpha?

$H_0: \mu_1 = \mu_2 = \mu_3$

H_A: at least one pair of population means differ

$\alpha = 0.05$

(c) What is the estimate of the within-group variance?

$$S_W^2 = \frac{(22-1)\,8.67^2 + (18-1)\,7.79^2 + (23-1)\,8^2}{(22+18+23-3)}$$

$$= \frac{4020}{60} = 67$$

(d) What is the estimate of the between-groups variance?

$$S_B^2 = \frac{(62.55-62.13)^2 \cdot 22 + (63.13-62.13)^2 \cdot 18 + (60.83-62.13)^2 \cdot 23}{3-1}$$

$$= \frac{66.6}{2} = 33.3$$

(e) What is the value of the test statistic?

test statistic $\rightarrow F = \frac{S_B^2}{S_W^2} = \frac{33.3}{67} = 0.5$

(f) What distribution does the test statistic follow?

distribution $\rightarrow F_{k-1,n-k} = F_{2,60}$

(g) What is the p-value for the test?

$p > 0.1$ (from table in book)

$p > \alpha \rightarrow$ fail to reject

(h) Draw a conclusion for the test.

There is not sufficient evidence to suggest a difference in mean age across the 3 centers.

To do this test in R, use the following code:

```
# ANOVA
cen = c(rep("A", 22), rep("B", 18), rep("C", 23)) #rename the centers A, B and C
dataframe = data.frame(data$age, cen)
fit <- aov(age ~ cen, data=dataframe)
summary(fit)

> summary(fit)
            Df Sum Sq Mean Sq F value Pr(>F)
cen          2   66.6  33.307  0.4971 0.6108
Residuals   60 4020.4  67.006
```

age	cen	
62	A	1
65	A	
	A	
⋮	⋮	
	B	2
	B	
	⋮	
	C	3
	C	

(i) Calculate the Bonferroni corrected Type I error.

$$\alpha^* = \frac{\alpha}{\binom{K}{2}} = \frac{0.05}{\binom{3}{2}} = \frac{0.05}{\frac{3!}{1!\,2!}} = \frac{0.05}{3} = 0.0167$$

Now let's suppose that we actually rejected the null hypothesis. State what your new conclusion would be.

At least one pair of population means differ.

Perform a hypothesis test comparing the means of centers 1 and 2. What type of test should you perform? Go through the steps of hypothesis testing and state your conclusion in terms of the problem.

2-sample t-test with equal variances

$H_0: \mu_1 = \mu_2$

$H_A: \mu_1 \neq \mu_2$

$\alpha = 0.05$

$$T = \frac{62.55 - 63.28}{\sqrt{\left(\frac{1}{22} + \frac{1}{18}\right)\frac{(22-1)8.67^2 + (18-1)7.79^2}{22+18-2}}} = -0.277$$

t_{38} distribution

p-value $= 0.39 > \alpha \longrightarrow$ fail to reject H_0

There is not sufficient evidence to suggest a difference in mean age between centers 1 and 2.

→ this shouldn't be surprising since we failed to reject the null hypothesis $\mu_1 = \mu_2 = \mu_3$.

Lecture 9: Nonparametric Methods

Supplementary Reading: Pagano/Gauvreau; Chapters 13

Motivation

For the statistical tests that we've studied up to this point, the populations from which the data were sampled were assumed to be either normally distributed or approximately so. In fact, this property is necessary for the tests to be valid. Since the forms of the underlying distributions are are assumed to be known and only the values of certain parameters are not, these tests are said to be parametric. If the data do not conform to the assumptions made by such traditional techniques, nonparametric methods of statistical inference should be used instead. *Nonparametric techniques* make fewer assumptions about the nature of the underlying distributions. As a result, they are sometimes called *distribution-free methods*.

Nonparametric Tests

These tests are considered *nonparametric* because they make no assumptions about the distribution of the data (as compared to parametric tests like the t-test, which assumes the data follow a particular distribution). If you "know" the distribution that generated your data, these tests are still valid, but not as powerful as their parametric counterparts.

The Sign Test

- Used with dependent (paired) data
- Similar to the paired t-test, but does not require the differences to be Normally distributed
- Null hypothesis: In the underlying population of differences among pairs, the median difference is equal to 0
- Test statistic: $z_+ = \dfrac{D - (n/2)}{\sqrt{n/4}}$ where
  - D = # of positive differences
  - n = # of nonzero differences
  - $z_+ \sim N(0,1)$ for $n >$ approximately 20, use $D \sim \text{Binomial}(n, 1/2)$ for $n <$ approximately 20
- Calculated in Stata using `signtest var1 = var2`.

Wilcoxon Signed-Rank Test

- Used with dependent (paired) data
- Similar to the paired t-test, but does not require the differences to be Normally distributed
- Unlike the sign test, takes into account the magnitude of the differences
- Null hypothesis: In the underlying population of differences among pairs, the median difference is equal to 0
- Test statistic: $z_T = \dfrac{T - n(n+1)/4}{\sqrt{\frac{n(n+1)(2n+1)}{24}}}$ where
  - T = the smaller of { sum of positive ranks, sum of negative ranks } (ignoring signs)
  - n = # of nonzero differences
  - $z_T \sim N(0,1)$ for $n >$ approximately 20, use tables (e.g., Table A.6 in textbook) when n is small
- Calculated in Stata using `signrank var1 = var2`.

Wilcoxon Rank Sum Test

- Also known as the Mann-Whitney-Wilcoxon test, the Wilcoxon-Mann-Whitney test, and the Mann-Whitney U test (all refer to the same test)
- Used with independent (unpaired) data
- Similar to the two-sample t-test, but does not require data from the two groups to be Normally distributed. But, does assume that the two distributions have the same general shape.
- Null hypothesis: The medians of the two populations are identical
- Test statistic: $z_W = \dfrac{W - n_s(n_s + n_L + 1)/2}{\sqrt{\frac{n_S n_L (n_S + n_L + 1)}{12}}}$ where
  - W = the smaller of { sum of ranks in group 1, sum of ranks in group 2 }
  - n_S = # of observations in the sample with the smaller sum of ranks
  - n_L = # of observations in the sample with the larger sum of ranks
  - $z_W \sim N(0, 1)$ for $n >$ approximately 20, use tables (e.g., Table A.7 in textbook) when n is small
- Calculated in Stata using `ranksum var1 = var2`, or using `ranksum var1, by(groupVar)`.

Example

A study was conducted to evaluate the effectiveness of a work site health promotion program in reducing the prevalence of cigarette smoking. Thirty-two work sites were randomly assigned either to implement the health program or to make no changes for a period of two years. The promotion program consisted of health education classes combined with a payroll-based incentive program. The data collected during the study are saved in the dataset `program.dta` and `program.csv` .

For each work site, smoking prevalence at the start of the study is saved under the variable `baseline`, and smoking prevalence at the end of the two-year period under the name `followup`. The variable `group` contains the value 1 for work sites that implemented the health program and 2 for sites that did not. The data are below:

group	baseline	followup	difference
2	16.50	18.02	-1.52
2	29.60	29.68	-0.08
2	24.80	19.27	5.53
2	31.11	27.35	3.76
2	26.65	23.70	2.95
2	16.66	17.73	-1.07
2	28.06	25.74	2.32
2	9.85	12.44	-2.59
2	20.37	15.64	4.73
2	26.66	28.76	-2.10
2	28.13	27.35	0.78
2	26.85	26.39	0.46
2	25.71	25.15	0.56
2	24.09	24.16	-0.07
2	23.25	25.13	-1.88
2	21.87	18.64	3.23
1	28.61	24.34	4.27
1	27.56	27.71	-0.15
1	32.21	22.15	10.06
1	25.22	21.33	3.89
1	26.44	23.76	2.68
1	28.93	28.93	0.00
1	22.26	16.39	5.87
1	29.55	26.15	3.40
1	22.67	19.70	2.97
1	25.78	19.54	6.24
1	15.41	13.49	1.92
1	28.03	28.47	-0.44
1	23.90	21.52	2.38
1	15.82	13.99	1.83
1	19.09	16.84	2.25
1	24.51	23.02	1.49

1. For the work sites that implemented the health promotion program, test the null hypothesis that the median difference in smoking prevalence over the two-year period is equal to 0.

- Using R:
  `wilcox.test(data$baseline[17:32], data$followup[17:32], paired=TRUE)`
  Note: we are using the data contained in rows 17-32 because those rows correspond to the work sites that implemented the health program (group 1).

```
> wilcox.test(data$baseline[17:32], data$followup[17:32], paired=TRUE)

Wilcoxon signed rank test with continuity
correction

data:  data$baseline[17:32] and data$followup[17:32]
V = 117, p-value = 0.001332
alternative hypothesis: true location shift is not equal to 0
```

p-value $< \alpha$ $\rightarrow$ reject H_0

Using the signed-rank test, the p-value is equal to 0.0013. We reject the null hypothesis that the median difference in smoking prevalence is equal to 0. For work sites that implemented the health promotion program, the median difference is greater than 0, suggesting that prevalence decreased over the two-year period.

(baseline) - (follow-up)

+ : lower prevalence at follow-up
0 : no difference
− : higher prevalence at follow-up

$H_0: m_{baseline} = m_{follow\text{-}up}$ m = median

$H_A: m_{baseline} \neq m_{follow\text{-}up}$

2. Test the same null hypothesis for the sites that did not make any changes.

   - In R: `wilcox.test(data$baseline[1:16], data$followup[1:16], paired=TRUE)`

```
> wilcox.test(data$baseline[1:16], data$followup[1:16], paired=TRUE)

Wilcoxon signed rank test

data:  data$baseline[1:16] and data$followup[1:16]
V = 92, p-value = 0.2312
alternative hypothesis: true location shift is not equal to 0
```

p-value > α → fail to reject

Again using the signed-rank test, the *p*-value is equal to 0.23. Therefore, we are unable to reject the null hypothesis that the median difference in smoking prevalence is equal to 0. For work sites that did not implement the health promotion program, there is no evidence that smoking prevalence changed over the two-year period.

3. Evaluate the null hypothesis that the median difference in smoking prevalence over the two-year period for work sites that implemented the program is equal to the median difference for sites that did not.

   - Using R:

```
> wilcox.test(data$difference[17:32], data$difference[1:16])

Wilcoxon rank sum test

data:  data$difference[17:32] and data$difference[1:16]
W = 76, p-value = 0.05134
alternative hypothesis: true location shift is not equal to 0
```

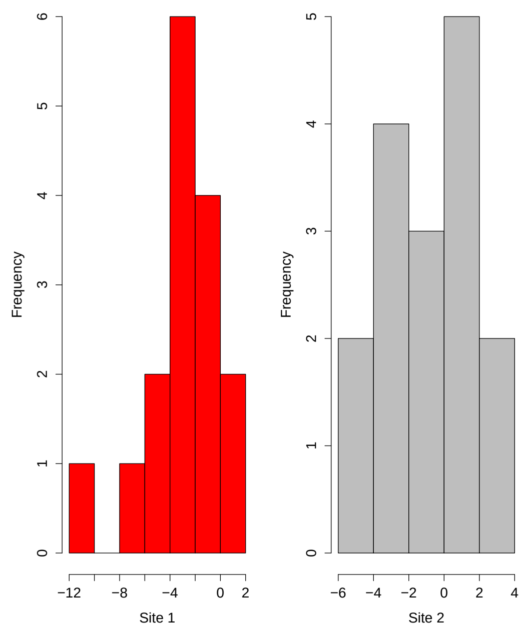
> 0.05

The rank sum test is used for this comparison, because we're comparing the differences in group 1 to the differences in group 2 – and these two groups of differences are considered independent. Since the *p*-value is equal to 0.05, we are on the borderline between rejecting and not rejecting the null hypothesis that the medians are identical. We conclude that there is some evidence that the medians are not equal, and that changes in the group that implemented the program tend to be larger than changes in the group that did not.

4. Could the two-sample t-test be used to analyze these data? Why or why not?

   In R:

```
# Histograms of differences
par(mfrow = c(1,2))
hist(data$difference[17:32], xlab = "Site 1", main = "", col = "red")
hist(data$difference[1:16], xlab = "Site 2", main = "", col = "gray")
```



   In a two-sample t-test, we would be comparing the mean differences in group 1 (the group that implemented the program) to the mean differences in group 2 (the group that did not implement the program). One of the assumptions of the t-test is that our sample size is large enough, which is the case here (32 observations). Another assumption is that the distribution of the variable (in the case, the distribution of the differences) is Normal within each group. In the histograms above, we see that this is not true for one of the groups. So the t-test would not be appropriate here.

5. Do you believe that the health promotion program was effective in reducing the prevalence of smoking? Explain.

Lecture 10: Proportions

Supplementary Reading: Pagano/Gauvreau; Chapter 14

Motivation

Example: Breast Cancer and Birth.
If you wait longer to have children, are you at increased risk for breast cancer?

Case/Control Study

- 3320 women with breast cancer (cases)
- 10,245 women without breast cancer (controls)
- Age at 1^{st} childbirth divided into two groups ($\leqslant 29, \geqslant 30$).
  - 683 (21.2%) cases had 1^{st} child at age $\geqslant 30$
  - 1498 (14.6%) controls had 1^{st} child at age$\geqslant 30$

Is this difference significant or due to chance?

Comparing two proportions

Two commonly used methods to compare two proportions of cases to controls:

1. Normal approximation to the binomial
2. Contingency tables

The binomial distribution provides the foundation for analyzing proportions.

Review: Binomial Distribution

- n independent Bernoulli trials [with 2 outcomes, say, flipping a coin]
- random variable (X) is the number of successes in n trials
- probability of success at each trial is the same (p)
- probability of observing exactly x successes in n trials is defined as
  $\mathrm{P}(X = x) = \left(\begin{array}{c} n \\ x \end{array} \right) p^x (1-p)^{n-x}$
- mean number of successes: $\mathrm{E}(X) = np$
- variance: $\mathrm{Var}(X) = np(1-p)$

Normal theory method

The Binomial distribution is cumbersome when n is large.

Approximate $\mathrm{P}(X = x)$ with normal distribution.

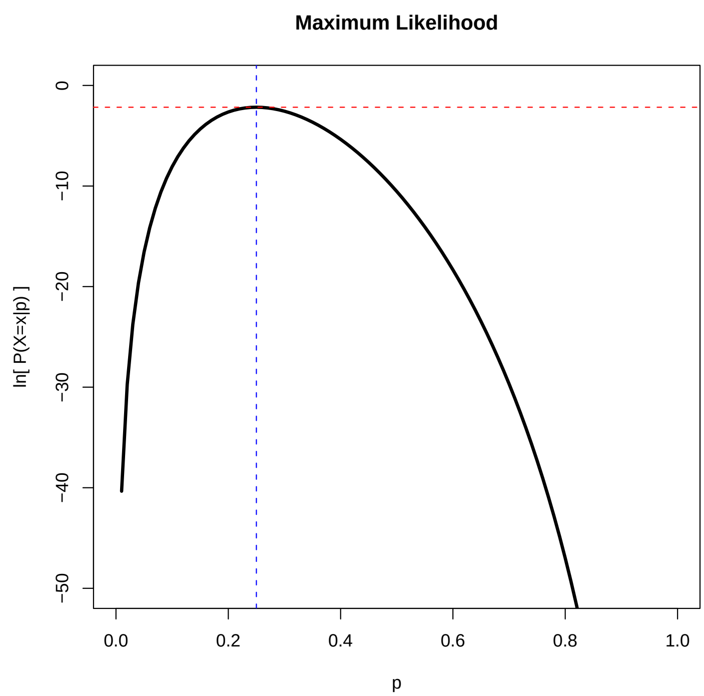
- When n is large (usually np & $n(1-p) \geqslant 5$)
- Test statistic: $Z = \frac{X-\mathrm{E}(X)}{\sqrt{\mathrm{var}(X)}} = \frac{X-np}{\sqrt{np(1-p)}} \;\sim\; \mathrm{N}(0,1)$
- Equivalent test statistic: $Z = \frac{\widehat{p}-p}{\sqrt{p(1-p)/n}} = \frac{\widehat{p}-p_0}{\sqrt{p_0(1-p_0)/n}} = \frac{X-np}{\sqrt{np(1-p)}}$

  Where p_0 is the null proportion value, or the value you are comparing your sample proportion to.

  - $\widehat{p} = X/n$ is <u>sample</u> proportion
  - p is <u>population</u> proportion

Sample proportion $\hat{p}$ has the following properties:

- It is the **maximum likelihood estimate** (MLE) of p - (the value of the parameter p that is most "likely" to have generated the observed sample)
- The mean of the sampling distribution of $\hat{p}$ is p
- The standard deviation of the distribution of $\hat{p}$'s is equal to $\sqrt{\hat{p}(1-\hat{p})/n}$
- The shape of sampling distribution of the distribution of $\hat{p}$s is approximately normal provided n is sufficiently large. (Why?)



Derivation of confidence intervals and hypothesis testing are done in a similar manner as was done with sample means.

- 95% CI for p: $\hat{p} \pm 1.96\sqrt{p(1-p)/n}$ where p is estimated by $\hat{p}$
- reject $H_0 : p = p_0$ when $|Z| \geqslant 1.96$ (2-sided test), or when p-value $< \alpha$

  Note: don't confuse the population parameter p with the p-value p.

Example: Lung cancer survival - pg. 329 in Pagano book

Suppose the 5-year survival for individuals under the age of 40 who have been diagnosed with lung cancer has unknown proportion p.

We do know the proportion of patients that survive for 5 years among those who are over 40 years of age at time of diagnosis is 8.2%.

Is it possible that this proportion is the same for those under 40? Suppose we sample 52 lung cancer patients who are under the age of 40, and find that 6 of them survive for 5 years. Perform a hypothesis test to see if it's plausible that this proportion is the same as the proportion of survivors in the over 40 age group.

- $p_0 =$ 0.082
- $\hat{p} =$ $6/52 = 0.115$
- $H_0 :$ $p = 0.082$
- $H_1 :$ $p \neq 0.082$
- $\alpha =$ 0.05
- Test statistic: $Z = \dfrac{\hat{p} - p_0}{\sqrt{\dfrac{\hat{p}(1-\hat{p})}{n}}} = \dfrac{0.115 - 0.082}{\sqrt{0.115(1-0.115)/52}} = 0.75$
- Reject or fail to reject? $P(Z \geq 0.75) = 0.23$
- Conclusion:

$p\text{-value} = 2 \cdot 0.23 = 0.46 > \alpha \rightarrow$ fail to reject H_0

This sample does not provide enough evidence to suggest the proportion of patients with lung cancer who have survived 5 years is different between the two age groups.

Comparing 2 proportions

In the breast cancer example we were interested in whether the proportion of women that were older than 30 for the birth of their first child is different for cases and controls.

Let

- n_1=number of cases
- n_2=number of controls
- p_1=prob age at 1^{st} birth is $\geqslant 30$ for cases
- $\widehat{p}_1$=sample proportion with age at 1^{st} birth $\geqslant 30$ for cases
- p_2=prob age at 1^{st} birth is $\geqslant 30$ for controls
- $\widehat{p}_2$=sample proportion with age at 1^{st} birth $\geqslant 30$ for controls

Do the same thing we did for differences in two means:

- Define hypotheses: $H_0 : p_1 = p_2$ vs. $H_1 : p_1 \neq p_2$
- Test statistic: $Z = \frac{\widehat{p}_1 - \widehat{p}_2}{\sqrt{p(1-p)(1/n_1+1/n_2)}} \sim \mathrm{N}(0, 1)$
- Select an α-level that will define the region in which H_0 will be rejected. With $\alpha = 0.05$, we reject if $|Z| \geqslant 1.96$ (2-sided test)
- Draw sample and compute $\widehat{p}_1$ and $\widehat{p}_2$
- What about p?
  Estimate with $\widehat{p} = \frac{n_1\widehat{p}_1 + n_2\widehat{p}_2}{n_1+n_2}$

  $\Rightarrow Z = \frac{\widehat{p}_1 - \widehat{p}_2}{\sqrt{\widehat{p}(1-\widehat{p})(1/n_1+1/n_2)}} \sim \mathrm{N}(0, 1)$

Summary of Tests

	One-Sample	Two-Sample
Hypotheses	$H_0 : p = p_0$ $H_A : p \neq p_0$	$H_0 : p_1 = p_2$ $H_A : p_1 \neq p_2$
Test Statistic	$Z = \frac{\hat{p}-p_0}{\sqrt{p_0(1-p_0)/n}}$	$Z = \frac{(\hat{p}_1-\hat{p}_2)-(p_1-p_2)}{\sqrt{\frac{n_1\hat{p}_1+n_2\hat{p}_2}{n_1+n_2}(1-\frac{n_1\hat{p}_1+n_2\hat{p}_2}{n_1+n_2})[(1/n_1)+(1/n_2)]}}$
Distribution of Test Statistic	standard normal	standard normal
Confidence Intervals	$(\hat{p} \pm z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}})$	$(\hat{p}_1 - \hat{p}_2 \pm z_{1-\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}})$

Example: Breast Cancer

- 3320 women with breast cancer (cases)
- 10,245 women without breast cancer (controls)
- Age at 1^{st} childbirth divided into two groups ($\leqslant 29, \geqslant 30$).
  - 683 (21.2%) cases had 1^{st} child at age $\geqslant 30$
  - 1498 (14.6%) controls had 1^{st} child at age $\geqslant 30$

Perform a test to see if there is a significant difference between the proportions of each group.

$H_0: p_1 = p_2$

$H_A: p_1 \neq p_2$

$\alpha = 0.05$

$n_1 = 3220$

$n_2 = 10{,}245$

$\hat{p}_1 = 0.212$

$\hat{p}_2 = 0.146$

$$\hat{p} = \frac{0.212(3220) + 0.146(10245)}{3220 + 10245}$$

$$= 0.162$$

$$Z = \frac{0.212 - 0.142}{\sqrt{0.162(1-0.162)\left(\frac{1}{3220} + \frac{1}{10245}\right)}} = 8.9$$

p-value < 0.05 = α → reject Ho

There is evidence to suggest women with breast cancer are more likely to have had their first child at or after the age of 30 than comparable women without breast cancer.

- In R:

```
> prop.test(x=c(683,1498), n=c(3220, 10245))

2-sample test for equality of proportions with
continuity correction

data:  c(683, 1498) out of c(3220, 10245)
X-squared = 77.8851, df = 1, p-value < 2.2e-16
alternative hypothesis: two.sided
95 percent confidence interval:
 0.04999981 0.08178846
sample estimates:
   prop 1    prop 2
0.2121118 0.1462177
```

Example: Cognitive Ability

Suppose we are interested in investigating the cognitive abilities of children weighing less than 1500 grams at birth. Although their birth weights are extremely low, many of these children exhibit normal growth patterns during the first year of life. A small group does not. These children suffer from perinatal growth failure, a condition that prevents them from developing properly. One indicator of perinatal growth failure is that during the first several months of life, the infant has a head circumference measurement that is far below normal.

We would like to examine the relationship between perinatal growth failure and subsequent cognitive ability. In particular, we wish to estimate the proportion of children, p, suffering from this condition who, when they reach 8 years of age, have intelligence quotient (IQ) scores that are below 70. In the general population, IQ scores are scaled to have mean 100; a score less than 70 suggests a deficiency in cognitive ability. To estimate the proportion of children with IQs in this range, a random sample of 33 infants with perinatal growth failure was chosen. At the age of 8, eight children have scores below 70.

For this problem, it will turn out that the normal distribution is not a good approximation to the sampling distribution of a sample proportion. For the purposes of illustration, however, the problem asks you to compute tests and confidence intervals using both the normal approximation and exact methods.

1. Find a point estimate for the population proportion p.

$$\hat{p} = \frac{x}{n} = \frac{8}{33} = 0.24$$

2. Use the normal approximation to construct a 95% confidence interval for the population proportion p.

$$\left(\hat{p} \pm Z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)$$

$$= \left(0.24 \pm 1.96 \sqrt{\frac{0.24(1-0.24)}{33}}\right)$$

$$= (0.0943, 0.3857)$$

3. Although we do not know the true value of p for this population, we do know that 3.2% of the children who exhibited normal growth in the perinatal period have IQ scores below 70 when they reach school age. We would like to know whether this is also true of the children who suffered from perinatal growth failure. Since we are concerned with deviations that could occur in either direction, conduct a two-sided test at the 0.05 level of significance. Although np_0 here is 1.056, use a normal approximation.

i) What are the null and alternative hypotheses? What is alpha?

$H_0: p = 0.032$

$H_A: p \neq 0.032$

$\alpha = 0.05$

ii) What is the value of the test statistic? What is the distribution of the test statistic?

$$Z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{0.24 - 0.032}{\sqrt{\frac{0.032(1-0.032)}{33}}} = 6.79$$

$Z \sim N(0,1)$ (standard Normal)

iii) What is the p-value of the test?

$6.79 > 1.96 \rightarrow$ p-value $<< 0.05 = \alpha$

iv) Do you reject or fail to reject the null hypothesis? What do you conclude?

p-value $< \alpha \rightarrow$ reject H_0

There is evidence to suggest the proportion of children who have low IQ scores among those with perinatal growth failure is different than the general population.

In R:

```
> prop.test(8, 33, 0.032)

1-sample proportions test with continuity
correction

data:  8 out of 33, null probability 0.032
X-squared = 40.623, df = 1, p-value = 1.846e-10
alternative hypothesis: true p is not equal to 0.032
95 percent confidence interval:
 0.1174329 0.4263056
sample estimates:
        p
0.2424242

Warning message:
In prop.test(8, 33, 0.032) : Chi-squared approximation may be incorrect
```

4. We can also perform an exact binomial test when the Normal distribution is not a good approximation.

In R:

```
> binom.test(x=8, n=33, p=0.032)

Exact binomial test

data:  8 and 33
number of successes = 8, number of trials = 33,
p-value = 7.445e-06
alternative hypothesis: true probability of success is not equal to 0.032
95 percent confidence interval:
 0.1109233 0.4225893
sample estimates:
probability of success
             0.2424242
```

i) What is the p-value of the test?

p-value ≈ 0.000007

ii) Do you reject or fail to reject the null hypothesis at the 0.01 level of significance? What do you conclude?

p-value < α → reject H_0

Same conclusion as before.

Lecture 11: Contingency Tables

Supplementary Reading: Pagano/Gauvreau; Chapter 15

Motivation

Suppose we have data on 793 people involved in accidents. We have that:

- Of the 793, 147 were wearing helmets.
- Of the 793, 646 were not wearing helmets.
- Among those wearing helmets, 17 suffered head injuries
- Among those not wearing helmets, 218 suffered head injuries

	Wearing Helmet		
Head Injury	Yes	No	Total
Yes	17	218	235
No	130	428	558
Total	147	646	793

Question: Is there an association between the incidence of head injury and the use of helmets among individuals involved in accidents?

We are interested in testing H_0: whether proportion of persons who had head injuries among those wearing helmets is equal to proportion of individuals who had head injuries among those not wearing helmets versus H_1: different proportions.

Test statistic:

- Independent samples: use Chi-square test
- Paired (dependent) samples: use McNemar's test

Compute p-value and compare to α-level or compare test statistic to critical value.

Chi-square test

Denote a table of observed counts as follows:

(data)

	Variable 2		
Variable 1	Yes	No	Total
Yes	a	b	a+b
No	c	d	c+d
Total	a+c	b+d	n

From this table we compute the expected counts, *assuming independence of row and column*, as the product of the row total and the column total divided by the total number of observations

Table of expected counts is,

	Variable 2		
Variable 1	Yes	No	Total
Yes	(a+b)(a+c)/n	(a+b)(b+d)/n	a+b
No	(c+d)(a+c)/n	(c+d)(b+d)/n	c+d
Total	a+c	b+d	n

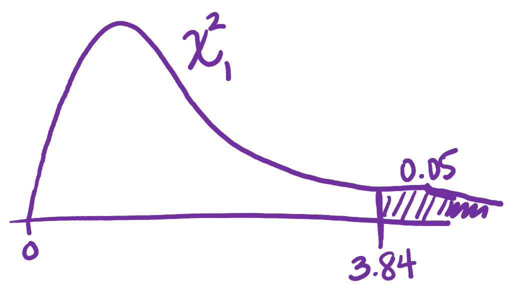
- Row/column totals are the same in observed and expected count tables because they are "fixed by design".
- If the observed table values are close to the expected table values we would conclude that there is no association. (fail to reject H_0)
- We use a chi-square test used to determine whether the deviations between observed and expected counts are too large to be attributed to chance.

$(1.96)^2 = 3.84$

Chi-square test statistic

test statistic

$$X^2 = \sum_{i=1}^{rc} \frac{(O_i - E_i)^2}{E_i}$$



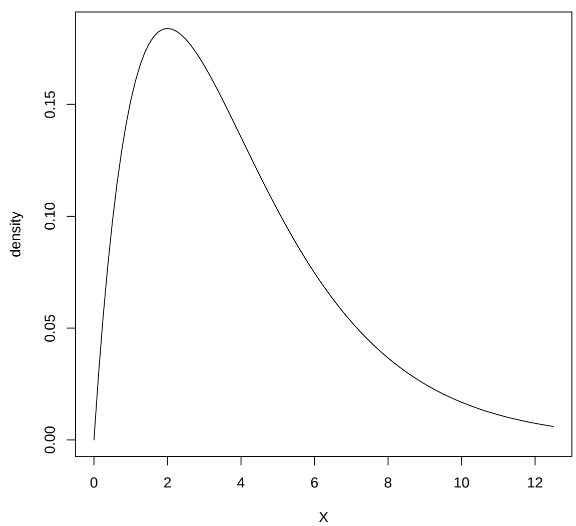
- r = number of rows; c =number of columns
- rc = number of cells in the table
- O_i = observed count for the i^{th} cell
- E_i = expected count for the i^{th} cell
- Probability distribution of X^2 is approximately a chi-squared distribution with $df = (r-1)(c-1)$ [denoted χ^2_{df}] distribution
- Reject if $X^2 > \chi^2_{df,\alpha}$ where $\chi^2_{df,\alpha}$ is the α-level critical value, i.e. $\mathrm{P}(X^2 > \chi^2_{df,\alpha}) = \alpha$
- Reject if p-value $< \alpha$

Properties of chi-squared distribution

- Not symmetric, skewed to the right
- A chi-square random variable can only take on positive values from 0 to ∞
- Distributions with small df are highly skewed.
- As df increases, the distribution becomes less skewed and more symmetric

- If $Z_1, ..., Z_{df} \sim \mathrm{N}(0,1)$ then $\sum_{i=1}^{df} Z_i^2 \sim \chi^2_{df}$.
- $\chi^2_{1,.05} = 3.84$ is equivalent to $Z^2_{.05} = 1.96^2$

Probability Density Function of a χ^2_4 Variable



Example: Breast cancer

– Table of observed counts:

	Age		
	≤ 29	$\geqslant 30$	Total
Cases	2537	683	3220
Controls	8747	1498	10245
Total	11284	2181	13465

– Table of expected counts:

	Age		
	≤ 29	$\geqslant 30$	Total
Cases	2698.4	521.6	3220
Controls	8585.6	1659.4	10245
Total	11284	2181	13465

Chi-square test statistic :

$$
\begin{aligned}
X^2 &= \sum_{i=1}^{rc} \frac{(O_i - E_i)^2}{E_i} \\
&= \frac{(2537 - 2698.4)^2}{2698.4} + \frac{(683 - 521.6)^2}{521.6} \\
&+ \frac{(8747 - 8585.6)^2}{8585.6} + \frac{(1498 - 1659.4)^2}{1659.4} \\
&= 78.29
\end{aligned}
$$

– $df = (r-1)(c-1) = (2-1)(2-1) = 1$

– Critical value is $\chi^2_{1,\underbrace{0.05}_{\alpha}} = 3.84$

– Critical region: reject H_0 if $X^2 > \chi^2_{1,.05} = 3.84$; $X^2 = 78.29 >> \overbrace{\chi^2_{1,.05}}^{3.84} \Rightarrow$ reject H_0

– Conclusion:

Breast cancer is significantly associated with having a first child after the age of 30.

– Note: this does not mean having a child over the age of 30 causes breast cancer. We are detecting association, NOT causality.

Note:

- For 2×2 tables, the X^2 test statistic has approximate chi-squared distribution with 1 df
- Equivalent to normal approximation of the binomial distribution
- Using discrete observations to estimate X^2
- Approximation may not be good when you have small df
- Apply what is called the **Yates correction** to get test statistic $X_C^2 = \sum_{i=1}^{4} \frac{(|O_i - E_i| - .5)^2}{E_i}$

Example: Breast Cancer

$$
\begin{aligned}
X_C^2 &= \frac{(|683 - 521.6| - .5)^2}{521.6} + \frac{(|2537 - 2698.4| - .5)^2}{2698.4} \\
&+ \frac{(|1498 - 1659.4| - .5)^2}{1659.4} + \frac{(|8747 - 8585.6| - .5)^2}{8585.6} \\
&= 77.81
\end{aligned}
$$

Note that the two test statistics are very close, and the conclusions are the same.

Point Estimates and Confidence Intervals

How do we quantify the strength of an association?

Odds ratio (OR):

- Odds in favor of an event that occurs with probability p are "$p/(1-p)$ to 1"
- If the odds in favor of event are a to b, then the probability event occurs is $a/(a+b)$
- If we have the following:

	Exposed	Unexposed	Total
Disease	a	b	a+b
No Disease	c	d	c+d
Total	a+c	b+d	n

- Odds ratio = odds in favor of disease among exposed individuals divided by odds in favor of disease among the unexposed
- Also called relative odds

Explicitly, the odds ratio is defined as:

$p\ /1-p$

$$OR = \frac{P(\text{disease} \mid \text{exposed})/(1-P(\text{disease} \mid \text{exposed}))}{P(\text{disease} \mid \text{unexposed})/(1-P(\text{disease} \mid \text{unexposed}))}$$
→ exposed group
→ unexposed group

Let D be the event that individual has the disease, and E be the event that the individual was exposed

- $P(D|E) = a/(a+c)$
- $P(\overline{D}|E) = 1 - a/(a+c) = c/(a+c)$ } odds exposed $= \frac{p}{1-p}$
- $P(D|\overline{E}) = b/(b+d)$
- $P(\overline{D}|\overline{E}) = 1 - b/(b+d) = d/(b+d)$ } $\text{odds}_{\text{unexposed}}$
- $\widehat{OR} = \frac{P(D|E)/P(\overline{D}|E)}{P(D|\overline{E})/P(\overline{D}|\overline{E})} = \frac{(a/(a+c))/(c/(a+c))}{(b/(b+d))/(d/(b+d))} = \frac{ad}{bc}$

H_0: - $OR = 1$ means identical odds (no association)

Interesting Relationship

Define the relative risk as:

$$\text{RR} = \frac{P(\text{disease}|\text{exposed})}{P(\text{disease}|\text{unexposed})}$$

In turns out that for rare diseases, the odds ratio is a close approximation of the relative risk. To see this, if we have:

$$P(\text{disease}|\text{exposed}) \approx 0$$

and

$$P(\text{disease}|\text{unexposed}) \approx 0$$

then

$$1 - P(\text{disease}|\text{exposed}) \approx 1$$

and

$$1 - P(\text{disease}|\text{unexposed}) \approx 1.$$

Therefore,

$$\begin{aligned} OR &= \frac{P(\text{disease} \mid \text{exposed})/(1 - P(\text{disease} \mid \text{exposed}))}{P(\text{disease} \mid \text{unexposed})/(1 - P(\text{disease} \mid \text{unexposed}))} \\ &\approx \frac{P(\text{disease}|\text{exposed})/1}{P(\text{disease}|\text{unexposed})/1} \\ &= \frac{P(\text{disease}|\text{exposed})}{P(\text{disease}|\text{unexposed})} \\ &= \text{RR} \end{aligned}$$

Example

Study to determine whether the use of electronic fetal monitoring (EFM) during labor affects the frequency of cesarean section deliveries

Data:

Cesarean	EFM exposure		
Section	Yes	No	Total
Yes	358	229	587
No	2492	2745	5237
Total	2850	2974	5824

- Odds of being delivered by C-section in the group that was monitored relative to the group that was not is

$$\widehat{OR} = \frac{ad}{bc} = \frac{358 \times 2745}{2492 \times 229} = 1.72$$

- Interpretation: the odds of being delivered by C-section for fetuses that are exposed to EFM during labor are 1.72 times greater than the odds for fetuses not exposed to EFM.
- Does not imply that EFM causes C-section delivery; it is possible fetuses that are monitored are at a higher risk.
- How good is this estimate?

To gauge uncertainty in this estimate, we compute confidence intervals:

- Recall 95% CI for mean μ is $(\bar{x} \pm 1.96\sigma/n)$
- Normality assumption not reasonable for odds ratio
- Probability distribution for OR is skewed to the right with values ranging from 0 to ∞
- Take natural logarithm of odds to get a distribution that is more symmetric and approximately normal

- 95% CI for $\ln(OR)$ is

$$(\ln(\widehat{OR}) - 1.96\ \text{SE}[\ln(\widehat{OR})], \ln(\widehat{OR}) + 1.96\ \text{SE}[\ln(\widehat{OR})])$$

- Where the estimated standard error of $\ln(\widehat{OR})$ is

$$\widehat{\text{SE}}[\ln(\widehat{OR})] = \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}$$

- 95% CI for OR is

$$(e^{\ln(\widehat{OR}) - 1.96\ \widehat{\text{SE}}[\ln(\widehat{OR})]}, e^{\ln(\widehat{OR}) + 1.96\ \widehat{\text{SE}}[\ln(\widehat{OR})]})$$

- To avoid problem of dividing by 0 when calculating the standard error, use the following when any of a, b, c, d are zero:

$$\widehat{\text{SE}}[\ln(\widehat{OR})] = \sqrt{\frac{1}{a+0.5} + \frac{1}{b+0.5} + \frac{1}{c+0.5} + \frac{1}{d+0.5}}$$

EFM Example:

- $\ln(\widehat{OR}) = \ln(1.72) = 0.542$
- $\widehat{\text{SE}}[\ln(\widehat{OR})] = \sqrt{\frac{1}{358} + \frac{1}{229} + \frac{1}{2492} + \frac{1}{2745}} = 0.89$
- 95% CI for natural logarithm of OR is
$(0.542 - 1.96\ \times 0.89, 0.542 + 1.96\ \times 0.89) = (0.368, 0.716)$
- Exponentiate to get CI for OR:$(e^{0.368}, e^{0.716}) = \underbrace{(1.44, 2.05)}_{1.72}$

H_0: OR = 1

null value (1) is not contained in the CI → reject H_0

Interpretation:

We are 95% confident that the odds of delivery by C-section among fetuses that are monitored are between 1.44 and 2.05 times the odds of fetuses that are not monitored.

The interval does not include 1 (which would mean that the fetuses that are monitored and those that are not monitored have identical odds of C-section delivery), meaning we reject H_0. If the interval did contain the value 1 (the null value), we would fail to reject the null.

Example

Consider the following data from a study that was performed at the Harvard School of Public Heath (HSPH) during the 2012 - 2013 academic year. All newly enrolled students were invited to participate in the study and were told that the study's aim was to investigate possible health effects of daily consumption of dark chocolate. The primary outcome of interest was cognitive ability at the end of the academic year, which was measured using the Wonderlic Personality Test. Among a sample of 70 students, 28 (the exposed group) ate dark chocolate at least once a week. Of the exposed students, 23 had above average cognitive ability scores at the end of the year. Of the 42 unexposed students, 26 had above average scores.

1. Complete the following 2 x 2 table:

	Outcome		
	Yes	No	Total
Exposed	23	5	28
Unexposed	26	16	42
Total	49	21	70

2. Perform a Chi-square test.
   i) What are the null and alternative hypotheses?

   H_0: no association between eating dark chocolate at least once a week and an above average test score.

   H_1: the is an association

   ii) Make a table of the expected counts.

$a = \frac{28 \cdot 49}{70} = 19.6$

$b = \frac{28 \cdot 21}{70} = 8.4$

$c = \frac{49 \cdot 42}{70} = 29.4$

$d = \frac{42 \cdot 21}{70} = 12.6$

19.6	8.4
29.4	12.6

Using R:

```
> data <- matrix(c(23, 5, 26, 16), nrow = 2, byrow = TRUE)
> colnames(data) <- c("Yes", "No")
> rownames(data) <- c("Exposed", "Unexposed")

> chisq.test(data)$expected         # prints the expected cell counts
           Yes   No
Exposed   19.6  8.4
Unexposed 29.4 12.6

> chisq.test(data)$observed         # prints the observed cell counts
          Yes No
Exposed    23  5
Unexposed  26 16
```

iii) What is the value and distribution of the test statistic? What is the p-value?

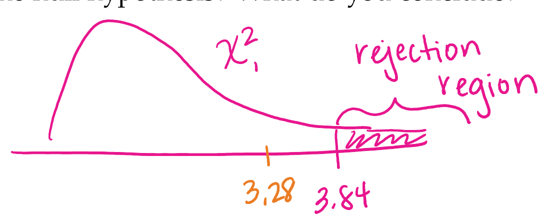
$$X^2 = \frac{(23-19.6)^2}{19.6} + \frac{(5-8.4)^2}{8.4} + \frac{(26-29.4)^2}{29.4} + \frac{(16-12.6)^2}{12.6} = 3.28$$

$$X^2 \overset{H_0}{\sim} \chi^2_1$$

iv) Do you reject or fail to reject the null hypothesis? What do you conclude?

$3.28 < 3.84$

$\rightarrow$ fail to reject H_0



v) Calculate $\widehat{OR}$:

$$\widehat{OR} = \frac{ad}{bc} = \frac{23(16)}{26(5)} = 2.83$$

$$95\%\ CI = (0.809,\ 11.313)$$

3. Now compare your results to the R output:

```
> chisq.test(data, correct=FALSE)    # performs the chi-squared test

Pearson's Chi-squared test

data:  data
X-squared = 3.2766, df = 1, p-value = 0.07027
```

no continuity correction

Matched Studies (Paired Data)

Example: Diabetic Retinopathy

Consider a study investigating a treatment for retinopathy among diabetics.

144 diabetics were treated in one (randomly selected) eye. The outcome of interest is whether the eyes progress to a serious stage of the disease.

Data:

	Treatment		
Progression	Yes	No	Total
Yes	46	25	71
No	98	119	217
Total	144	144	288

Is the proportion of progression in treated eyes the same as the proportion in untreated eyes?

- Have a total of 288 observations but only 144 pairs.
- Each individual provides 2 responses: one for the treated eye and one for the untreated eye. The eyes are *matched pairs*.

Can we use the chi-square test?

No, because it ignores the pairing in the data. Similar idea to the paired t-test versus the two-independent sample t-test.

Try to take this pairing into account.

Classify the data as follows:

	Treated Eye		
Untreated Eye	Progression	No Progression	Total
Progression	9	37	46
No Progression	16	82	98
Total	25	119	144

Note:

- Each entry in the table corresponds to the response of a pair of eyes (from an individual person), not to each individual eye.
- Of the 46 untreated eyes that progressed, 9 were paired with treated eyes that progressed and 37 were paired with treated eyes that did not progress.
- Of 98 untreated eyes that did not progress, 16 were paired with treated eyes that progressed and 82 were paired with treated eyes that did not progress.

Note:

- Pairs of data that provide information are those with (1) one treated/non-progressing matched with one untreated/ progressing and (2) one treated/progressing matched with one untreated/non-progressing
- These pairs are called **discordant** pairs

- These pairs correspond to number of pairs in the (1,2) cell (denoted r) and the (2,1) cell (denoted s) of the table - the off diagonal cells
- Ignore **concordant** pairs (those in (1,1) and (2,2) cells)
- Have to change H_0 from testing for equal probability of progression among treated and untreated eyes to testing for equal probability of each type of discordant pair.

McNemar's Test:

To conduct paired analysis, use McNemar's test.
H_0: "There is no association between treatment and progression."

- Test statistic: $X_M^2 = \frac{(|r-s|-1)^2}{r+s}$
- r is the the number of pairs in the (1,2) cell of table
- s is the number of pairs in the (2,1) cell
- X_M^2 is approximately distributed as χ_1^2

	r
s	

Back to the eye example:

	Treated Eye		
Untreated Eye	Progression	No Progression	Total
Progression	9	37	46
No Progression	16	82	98
Total	25	119	144

From the table we have:

- $r =$ 37
- $s =$ 16
- $X_M^2 = \frac{(|37-16|-1)^2}{37+16} = 7.55$
- $\alpha =$ 0.05
- $\chi^2_{1,.05} =$ 3.84
- $p-$value = 0.006 $< \alpha$
- Reject H_0 or fail to reject H_0? reject H_0
- Conclusion:

Given a population of diabetic retinopathy patients, treated eyes are less likely to progress to a severe stage of the disease than untreated eyes.

In R:

```
> data <- matrix(c(9, 37, 16, 82), nrow = 2, byrow = TRUE)
> colnames(data) <- c("Progression", "No Progression")
> rownames(data) <- c("Progression", "No Progression")
>
> mcnemar.test(data)

McNemar's Chi-squared test with continuity correction

data:  data
McNemar's chi-squared = 7.5472, df = 1, p-value = 0.00601
```