

Quantitative Methods

Serena DeStefani – Lecture 17 –8/3/2020

Announcements

- HW7 due tomorrow
- HW8 due Thursday
- Today: paired data + Chi-square tests
- Inference for regression
- Analysis of Variance (ANOVA)
- Careers
- → general review (start on Mon and Tue?)
- Next Wednesday: Final exam

Review

Inference about?	One sample or two?	Procedure	Model	Parameter	Estimate	SE	Chapter
Proportions	One sample	1-Proportion z -Interval	z	p	\hat{p}	$\sqrt{\frac{\hat{p}\hat{q}}{n}}$	19
		1-Proportion z -Test				$\sqrt{\frac{p_0q_0}{n}}$	20, 21
	Two independent groups	2-Proportion z -Interval	z	$p_1 - p_2$	$\hat{p}_1 - \hat{p}_2$	$\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}$	22
		2-Proportion z -Test				$\sqrt{\frac{\hat{p}\hat{q}}{n_1} + \frac{\hat{p}\hat{q}}{n_2}}, \hat{p} = \frac{y_1 + y_2}{n_1 + n_2}$	22
Means	One sample	t -Interval t -Test	t $df = n - 1$	μ	\bar{y}	$\frac{s}{\sqrt{n}}$	23
	Two independent groups	2-Sample t -Test 2-Sample t -Interval	t df from technology	$\mu_1 - \mu_2$	$\bar{y}_1 - \bar{y}_2$	$\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$	24
	n Matched pairs	Paired t -Test Paired t -Interval	t $df = n - 1$	μ_d	\bar{d}	$\frac{s_d}{\sqrt{n}}$	25

Chapter 23

Paired Samples and Blocks

23.1

Paired Data

Pairing

- Speed-skating races run in

pairs. One starts in the inner lane, the other in the outer lane.

- Above are some of the randomly assigned pairs.
- The data are **paired** rather than independent.
- **Blocking** involves pairing arising from an experiment.
- **Matching** involves pairing arising from an observational study.
- With pairing, we look at the **differences**.

Inner Lane		Outer Lane	
Name	Time	Name	Time
ZHANG Xiaolei	125.75	NEMOTO Nami	122.34
ABRAMOVA Yekaterina	121.63	LAMB Maria	122.12
REMPEL Shannon	122.24	NOH Seon Yeong	123.35
LEE Ju-Youn	120.85	TIMMER Marianne	120.45
ROKITA Anna Natalia	122.19	MARRA Adelia	123.07
YAKSHINA Valentina	122.15	OPITZ Lucille	122.75

Comparing Mileage

Name	5-Day Mileage	4-Day Mileage
Jeff	2798	2914
Betty	7724	6112
Roger	7505	6177
Tom	838	1102

Do flexible schedules reduce total mileage? The table shows some workers' mileage before and after the company switched to a 4-day work week.

- Why are these data paired?
 - Each driver's mileage recorded before and after.
 - Not independent since we expect a high “before” mileage to predict a high “after” mileage.
 - Paired because the differences such as Jeff: $2798 - 2914 = -116$ have meaning.

Differences for Speed-Skater Pairs

- For paired data, create a new data set of the differences.
- We can now look only at the differences.
- Ignoring the original data, we now have a single data set.
- Proceed with a one-sample t -test. This process is called a **paired t -test**.

Skating Pair	Inner Time	Outer Time	Inner – Outer
1	125.75	122.34	3.41
2	121.63	122.12	–0.49
3	122.24	123.35	–1.11
4	120.85	120.45	0.40
5	122.19	123.07	–0.88
6	122.15	122.75	–0.60

23.2

Assumptions and Conditions

Assumptions and Conditions

Paired Data Condition

- The data must be paired.
- Only use pairing if there is a natural matching.
- The two-sample t -test and the paired t -test are not interchangeable.

Independence Assumption

- For paired data, the groups are never independent.
- Need differences independent, not individuals
- Randomization ensures independence.

Assumptions and Conditions (Continued)

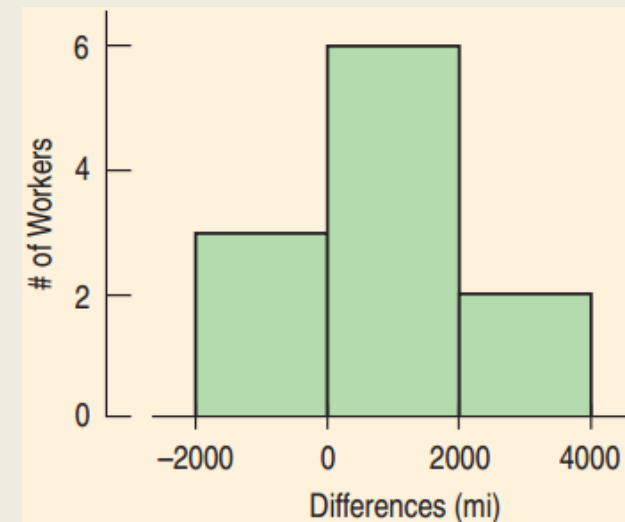
Normal Population Assumption

- Need to assume the differences follow a Normal model.
- **Nearly Normal Condition:**
 - Sketch a histogram and normal probability plot of the *differences*.
 - Normality less important for larger sample sizes.
 - Even if the individual measurements are skewed, bimodal or have outliers, the differences may still be Normal.

4-Day vs. 5-Day Mileage

Is it okay to use these data to test whether the new schedule changed the amount of driving?

- ✓ **Paired Data Condition:** Before and after study
- ✓ **Independence Assumption:** The driving behavior of one worker is independent of another.
- ✓ **Randomization Condition:** The work trips randomly occurred.
- ✓ **Nearly Normal Condition:** The sample size is small, but the histogram is unimodal and symmetric.
 - The **paired t -test** can be used.



The Paired t -Test

When the conditions are met, we can test whether the mean differences significantly differ from 0.

- $H_0: \mu_d = \Delta_0$ (Δ_0 is usually 0)
- $t = \frac{\bar{d} - \Delta_0}{SE(\bar{d})} \quad SE(\bar{d}) = \frac{s_d}{\sqrt{n}}$
- \bar{d} and s_d are the mean and standard deviation of the pairwise differences and n is the number of pairs.
- Use the Student's t -model with $n - 1$ degrees of freedom and find the P-value.

Speed-Skating Comparisons



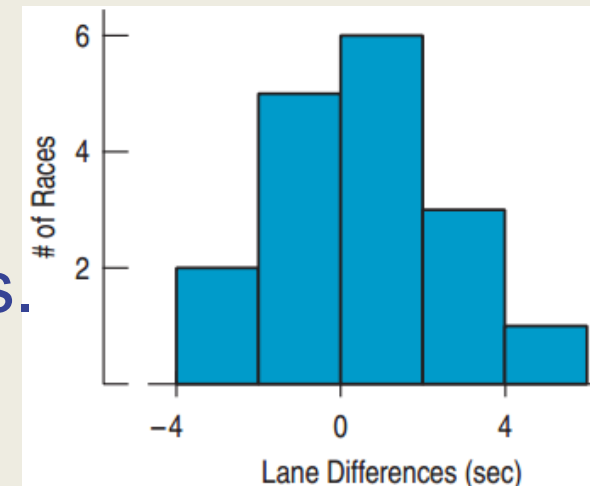
Was there a difference between speeds between inner and outer speed-skating lanes?

- **Plan:** I have data for 17 pairs of racers.
- **Hypotheses:**
 - $H_0: \mu_d = 0$
 - $H_A: \mu_d \neq 0$

Speed-Skating Comparisons



- **Model:**
 - ✓ **Paired Data Condition:** The racers compete in pairs.
 - ✓ **Independence Assumption:** Each race was independent of the others. Racers were assigned to lanes randomly.
 - ✓ **Nearly Normal Condition:** The histogram of the differences is unimodal and symmetric. No outliers.
- Use Student's t -model with 16 df.
- Perform a **paired t -test**.



Speed-Skating Comparisons



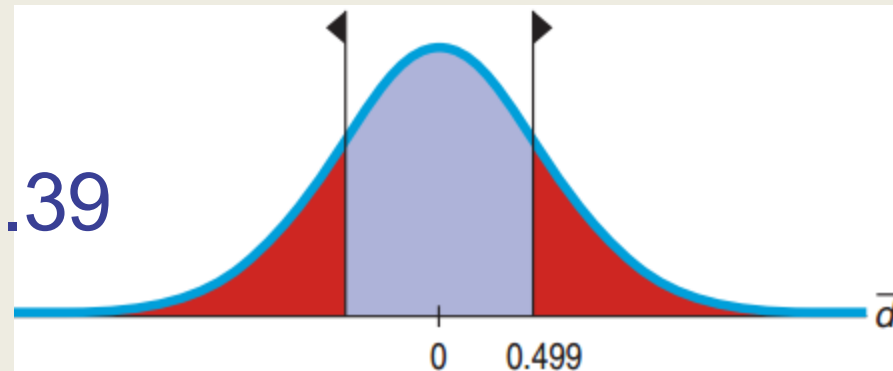
- **Mechanics:**

$n = 17$ pairs, $\bar{d} = 0.499$ sec., $s_d = 2.333$ sec.

$$SE(\bar{d}) = \frac{s_d}{\sqrt{n}} = \frac{2.333}{\sqrt{17}} = 0.5658$$

$$t_{16} = \frac{\bar{d} - 0}{SE(\bar{d})} = \frac{0.499}{0.5658} \approx 0.882$$

$$P\text{-value} = 2P(t_{16} > 0.882) = 0.39$$



Speed-Skating Comparisons



- **Conclusion:**
 - Events that occur more than a third of the time ($P\text{-value} = 0.39$) are not remarkable.
 - I can't conclude that the observed difference isn't due simply to random chance. It appears the fans may have interpreted a random fluctuation in the data as favoring one lane. There's insufficient evidence to declare any lack of fairness.

4-Day vs. 5-Day Mileage

Is there evidence that a 4-day work week vs. a 5-day work week would change how many miles workers drive?

- $H_0: \mu_d = 0$
- $H_A: \mu_d \neq 0$
- $n = 11$ pairs, $\bar{d} = 982$ miles, $s_d = 1139.6$ miles
- The assumptions and conditions are met. Use Student's t -model with 10 df.

4-Day vs. 5-Day Mileage

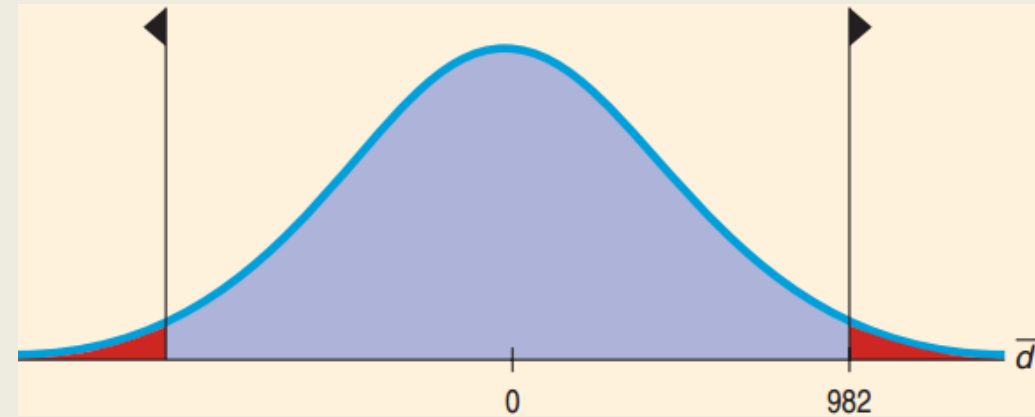
Is there evidence that a 4-day work week vs. a 5-day work week would change how many miles workers drive?

- $H_0: \mu_d = 0$
- $H_A: \mu_d \neq 0$
- $n = 11$ pairs, $\bar{d} = 982$ miles, $s_d = 1139.6$ miles
- The assumptions and conditions are met. Use Student's t -model with 10 df.

$$SE(\bar{d}) = \frac{s_d}{\sqrt{n}} = \frac{1139.6}{\sqrt{11}} = 343.6 \quad t_{10} = \frac{\bar{d} - 0}{SE(\bar{d})} = \frac{982.0}{343.6} \approx 2.86$$

4-Day vs. 5-Day Mileage

- P-value = $2P(t_{10} > 2.86) = 0.017$
- Since the P-value is small, reject H_0 .
- Conclude that the change in workweek did lead to a change in average driving mileage. It appears that changing the work schedule may reduce the mileage driven by workers.
- The confidence interval will help determine if there is practical significance.



23.3

Confidence Intervals for Matched Pairs

How Much Older is the Husband?

Wife's Age	Husband's Age	Difference (husband – wife)
43	49	6
28	25	–3
30	40	10
⋮	⋮	⋮

170 couples selected at random provided their ages.

- Goal: Find a confidence interval for the mean difference.

Paired t -interval

When the conditions are met, the confidence interval for the mean paired difference is

$$\bar{d} \pm t_{n-1}^* \times SE(\bar{d}) \qquad SE(\bar{d}) = \frac{s_d}{\sqrt{n}}$$

- The critical value t^* from the Student's t -model depends on the confidence level C and $df = n - 1$. (n is the number of pairs.)

How Much Older is the Husband?



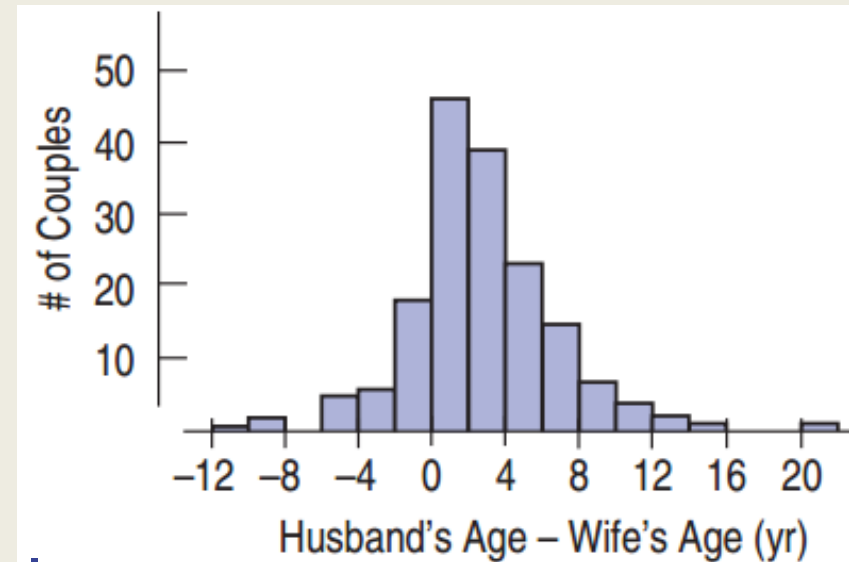
How big a difference is there, on average, between husbands and wives?

- **Plan:** I want the mean difference in age. Have a random sample of 200 couples with 170 providing ages their ages. $n = 170$
- **Model:**
 - ✓ **Paired Data Condition:** Married couples make a natural pairing.
 - ✓ **Independence Assumption:** The survey was randomized.

How Much Older is the Husband?



- **Model:**
 - ✓ **Nearly Normal Condition:**
The histogram is unimodal and symmetric.
- The conditions are met, so I can use the Student's t -model with $n - 1 = 169$ degrees of freedom and find a **paired t -interval**.



How Much Older is the Husband?



- **Mechanics:** $n = 170$, $\bar{d} = 2.2$ years, $s_d = 4.1$ years

How Much Older is the Husband?



- **Mechanics:** $n = 170$, $\bar{d} = 2.2$ years, $s_d = 4.1$ years

$$SE(\bar{d}) = \frac{s_d}{\sqrt{n}} = \frac{4.1}{\sqrt{170}} = 0.31 \text{ year}$$

$$t_{169} = 1.97 \text{ (from computer)}$$

$$ME = t_{169}^* \times SE(\bar{d}) = 1.97 \times 0.31 \approx 0.61$$

- The 95% CI for the difference between the means is
 $2.2 \pm 0.61 = 1.6 \text{ to } 2.8 \text{ years.}$

How Much Older is the Husband?



- **Conclusion:**

I am 95% confident that husbands are, on average, 1.6 to 2.8 years older than their wives.

Effect Size and the Work Week

We found a statistically significant difference between the driving mileage for the 4-day and 5-day work week. Is there a practical difference?

Effect Size and the Work Week

We found a statistically significant difference between the driving mileage for the 4-day and 5-day work week. Is there a practical difference?

- $\bar{d} = 982$ mi, $SE(\bar{d}) = 343.6$, $t_{10}^* = 2.228$ (for 95%)
- $ME = t_{10}^* \times SE(\bar{d}) = 2.228(343.6) = 765.54$
- 95% CI for μ_d is $982 \pm 765.54 = (216.46, 1747.54)$
- With 95% confidence, switching to a 4-day work week, I estimate the average mileage is between 216 and 1748 fewer miles per year. With high gas prices, this could save a lot of money.

23.4

Blocking

Disadvantages and Advantages of Pairing

Disadvantage

- Fewer degrees of freedom. Each pair considered as a single data value instead of two values.

Advantage

- Can significantly reduce variability by focusing just on what is being compared.
- Pairing is an example of effective blocking.

The advantage outweighs the disadvantage, so consider pairing if possible.

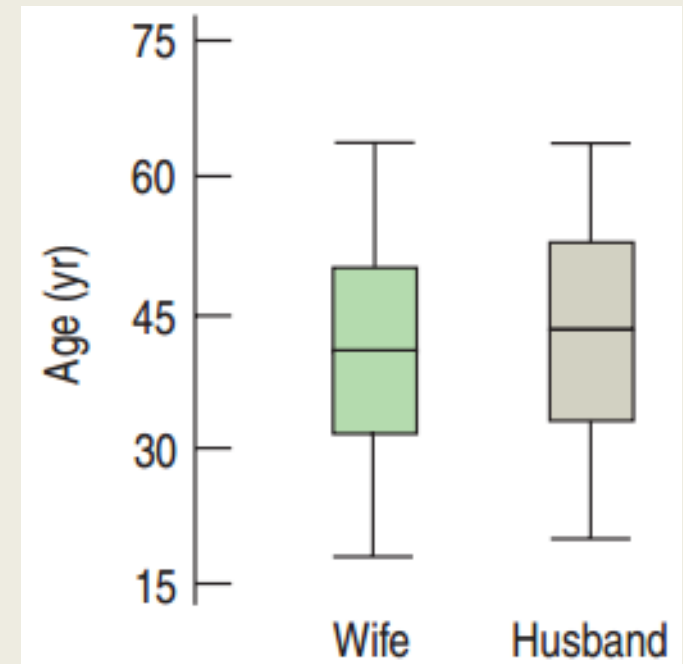
Comparing Pairs and No Pairs

Pairing

- With pairing husbands and wives, there was a clear difference in ages. **95% CI: (1.6, 2.8).**

Without Pairing

- Without pairing, this would not have been as conclusive.
- From the box plots the difference is hard to detect.



Dependence and Independence

- Previous methods required independence both between and within groups.
- With pairing, the samples are always *dependent*.
- With pairing, we need the pairs to be *independent* of each other.
- **Paired methods** are sometimes called methods for “dependent” samples.

What Can Go Wrong?

Don't use a two-sample t -test when you have paired data.

- Two-sample t -test and paired t -tests are not the same.

Don't use a paired t method when the samples aren't paired.

- Always check for a natural pairing such as husband/wife, before/after, younger sister/older sister.

What Can Go Wrong?

Don't forget outliers.

- Outliers of the differences not individuals. Make a plot of the differences to check for outliers.

Don't look for the difference between the means of paired groups with side-by-side box plots.

- The box plots of each group still contain the variation that pairing can remove. Comparing box plots is likely to be misleading.

Chapter 24

Comparing Counts

24.1

Goodness-of-Fit Test

Are CEO Zodiac Signs Uniform?



- A survey of **256** randomly selected CEOs
Zodiac signs
- If uniform, each sign expected
to have $256/12 \approx 21.3$ births
- Pisces has more.
- Others have fewer.
- Is the distribution far enough from uniform to
conclude the data do not come from a uniform
distribution?

Births	Sign	Births	Sign
23	Aries	18	Libra
20	Taurus	21	Scorpio
18	Gemini	19	Sagittarius
23	Cancer	22	Capricorn
20	Leo	24	Aquarius
19	Virgo	29	Pisces

Goodness-of-Fit vs. Two-Proportion t -Test

- To just test for Pisces, the two-proportion t -test works.
- We would need to perform 12 separate t -tests each with $\alpha = 0.05$, $P(\text{Type I Error})$ compounded...?
- Instead, the **Goodness-of-Fit Test** properly combines differences from expected into a single hypothesis test.
- For this test I need expected counts

Finding Expected Counts

1478 baseball players' birth months.

- Find the expected count for each month.

Month	Ballplayer Count	National Birth %	Month	Ballplayer Count	National Birth %
1	137	8%	7	102	9%
2	121	7%	8	165	9%
3	116	8%	9	134	9%
4	121	8%	10	115	9%
5	126	8%	11	105	8%
6	114	8%	12	122	9%
Total			Total	1478	100%

- January: $1478 \times 0.08 = 118.24$
- February: $1478 \times 0.07 = 103.46$

Month	Expected	Month	Expected
1	118.24	7	133.02
2	103.46	8	133.02
3	118.24	9	133.02
4	118.24	10	133.02
5	118.24	11	118.24
6	118.24	12	133.02

Assumptions and Conditions

Counted Data Condition

- The values in each cell are counts.
- Doesn't work with percents, proportions, or measurements

Independence Assumption

- The counts in each cell must be independent of each other.
- For random samples, we can generalize to the entire population.

Assumptions and Conditions (Continued)

Sample Size Assumption

- Expected counts for each cell ≥ 5 .
- This is called the **Expected Cell Frequency Condition**.

If the assumptions and conditions are met, we can perform a **Chi-Square Test for Goodness-of-Fit**.

Checking Assumptions and Conditions

- Are the assumptions and conditions met for the baseball player births study?
 - ✓ **Counted Data Condition:** Each cell is a count.
 - ✓ **Independence Assumption:** The births are independent. They do not come from a random sample, but they are representative of all players past and present.
 - ✓ **Expected Cell Frequency Condition:** Expected counts range from 103.46 to 133.02. All ≥ 5 .
- Goodness-of-fit does apply.

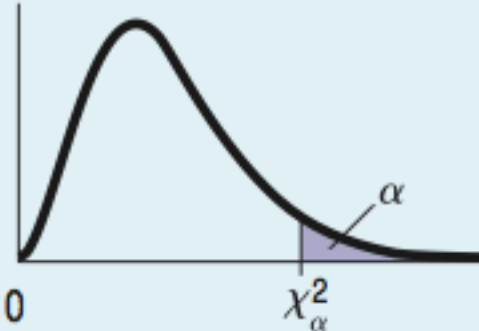
Chi-Square Calculations

- Interested in difference between observed and expected: **residuals**.
- Make positive by squaring them all.
- Get relative sizes of the residuals by dividing them by the expected counts.

$$\chi^2 = \sum \frac{(Obs - Exp)^2}{Exp}$$

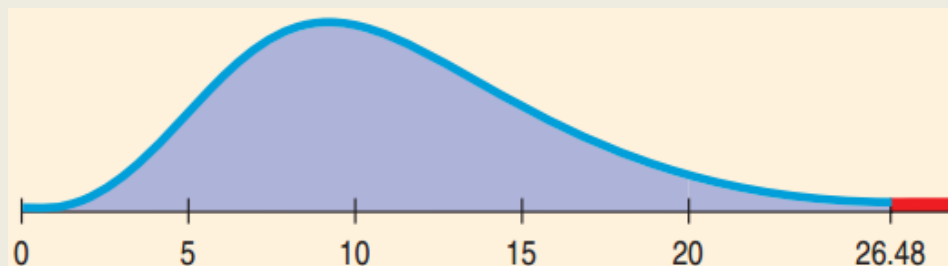
- This is a **Chi-Square Model** with $df = n - 1$.
- n is the number of categories, not the sample size.

Chi-Square P-Values

Right-Tail Probability		0.10	0.05	0.025	0.01	0.005
<div>  <p>Values of χ^2_α</p> </div>	df					
	1	2.706	3.841	5.024	6.635	7.879
	2	4.605	5.991	7.378	9.210	10.597
	3	6.251	7.815	9.348	11.345	12.838
	4	7.779	9.488	11.143	13.277	14.860
	5	9.236	11.070	12.833	15.086	16.750
	6	10.645	12.592	14.449	16.812	18.548
	7	12.017	14.067	16.013	18.475	20.278
	8	13.362	15.507	17.535	20.090	21.955
	9	14.684	16.919	19.023	21.666	23.589
	10	15.987	18.307	20.483	23.209	25.188
	11	17.275	19.675	21.920	24.725	26.757
	12	18.549	21.026	23.337	26.217	28.300
	13	19.812	23.362	24.736	27.688	29.819
	14	21.064	23.685	26.119	29.141	31.319

Baseball Birth Months

- H_0 : The distribution of baseball player birth months is the same as for the general population.
- H_A : The distribution of baseball player birth months is not the same as for the general population.



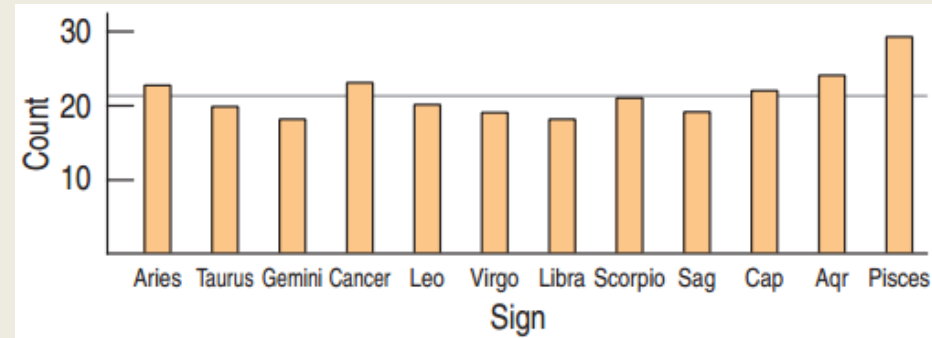
- $\chi^2_{11} = 26.48$ P-value = 0.0055 (or $0.01 < p < 0.005$)
- There's evidence that major league ballplayers' birth months have a different distribution from the rest of us.

Zodiac Signs of CEOs



- Plan: I have the zodiac signs of 256 CEOs. Is there a uniform distribution?
- Hypotheses:
 - H_0 : Births are uniformly distributed over zodiac signs.
 - H_A : Births are not uniformly distributed over zodiac signs.

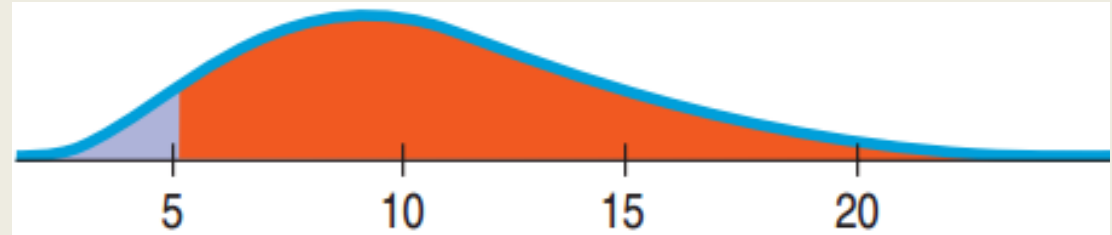
Zodiac Signs for CEOs



Model

- **Counted Data Condition:** I have counts of each sign.
- **Independence Assumption:** This is a convenience sample, but no reason to expect bias with respect to birthdays.
- **Expected Cell Frequency Condition:** $21.3 \geq 5$.
- The conditions are satisfied. I will use the **Chi-Square Goodness-of-Fit Model** with $12 - 1 = 11$ degrees of freedom.

Zodiac Signs for CEOs

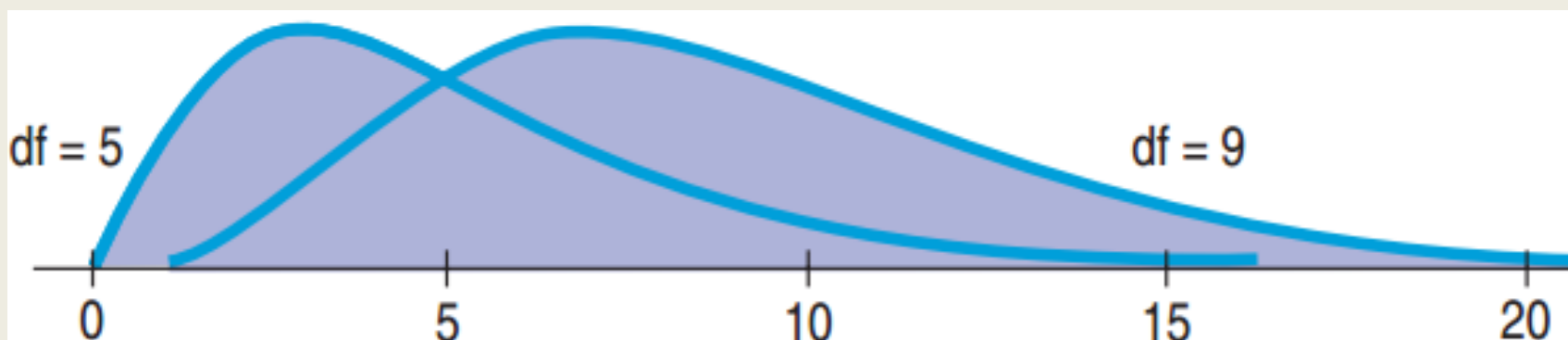


- **Mechanics:**

$$\chi_{11}^2 = 5.094, \text{ P-value} = 0.926 \text{ (from computer)}$$

- **Conclusion:** If the zodiac signs of CEOs were distributed uniformly, an observed chi-square value of 5.09 or higher would occur about 93% of the time.
- Fail to reject the H_0 . Conclude that there is no evidence of nonuniform distribution of zodiac signs among CEOs.

Why so big? The Shape of χ^2



- With **df** large (more cells), the weighted residuals add up quickly \rightarrow numerator gets bigger
- Unlike **z** or **t**, a larger χ^2 is more common. It all depends on the df: look at value 10
- The **mode** of χ^2 is **df - 2**.
- The **expected value** of χ^2 is **df**, to the right of the mode due to the right skewed distribution.
- The CEO curve had df=11, peaks at 9 and mean=11

$$\chi^2 = \sum \frac{(\text{Obs} - \text{Exp})^2}{\text{Exp}}$$

24.2

Chi-Square Test for Homogeneity

Activities of Graduates at Different Colleges

	Agriculture	Arts & Sciences	Engineering	ILR	Total
Employed	209	198	177	101	685
Grad School	104	171	158	33	466
Other	135	115	39	16	305
Total	448	484	374	150	1456

- Are graduates' activities the same at different colleges within the same university?
- Class-size differences cloud the table. Proportions may be helpful.

	Agriculture	Arts & Sciences	Engineering	ILR	Total
Employed	46.7%	40.9%	47.3%	67.3%	47.0%
Grad School	23.2%	35.3%	42.2%	22.0%	32.0%
Other	30.1%	23.8%	10.4%	10.7%	20.9%
Total	100.0%	100.0%	100.0%	100.0%	100.0%

Test for Homogeneity

- More than two proportions to compare, so we cannot use the two-proportion z-test.
- Generalize to the **chi-square test for homogeneity**.
- Same mechanics as goodness-of-fit, different hypotheses and conclusions.
- Chi-square test for homogeneity asks if the distribution is the same for different groups.
- The test looks for differences large enough to step beyond what we might expect from random sample-to-sample variation

Assumptions and Conditions

Same as Goodness-of-Fit

- **Counted Data Condition:** Always use counts, not proportions, percents, or measurements.
- **Independence Assumption:** Need randomization to generalize to the population.
- **Expected Cell Frequency Condition:** All expected counts should be at least 5.

Calculations

	Agriculture	Arts & Sciences	Engineering	ILR	Total
Employed	210.769	227.706	175.955	70.570	685
Grad School	143.385	154.907	119.701	48.008	466
Other	93.846	101.387	78.345	31.422	305
Total	448	484	374	150	1456

Expected Counts?

- 685, or about 47.0%, of the 1456 students who responded to the survey were employed.
- Of the 448 agriculture students, expect 47% of 448 or 210.76 to be employed.

χ^2 Calculation

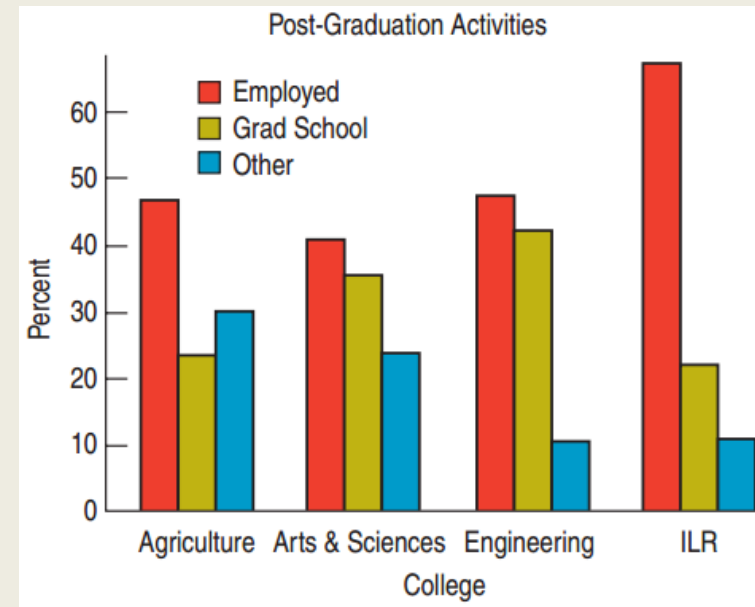
- Use the same formula with these observed and expected. $df = (R - 1)(C - 1) = (3 - 1)(4 - 1) = 6$.

Students' Choices Same Across Colleges?

- **Plan:** I have a table of counts from the college's class of 2011.
- **Hypotheses:**
 - H_0 : Students' post-graduation activities are distributed in the same way for all four colleges.
 - H_A : Students' plans do not have the same distribution.

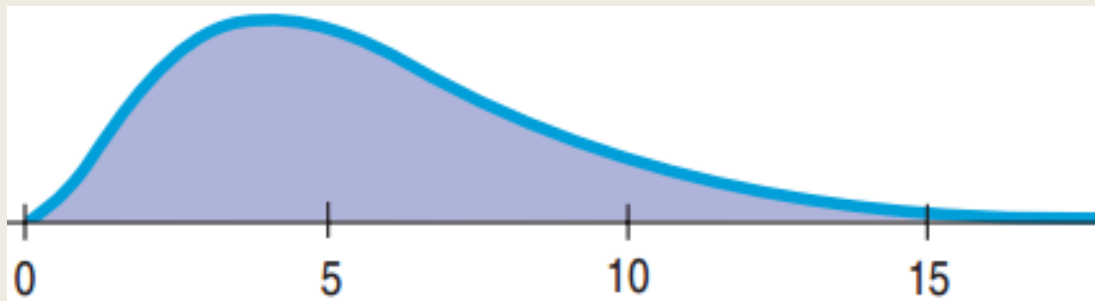
Students' Choices Same Across Colleges?

- **Model:** The bar chart shows how the percents differ.
- ✓ **Counted Data Condition:**
I will use the table of counts.
- ✓ **Independence Assumption:**
It is not a random sample, but with this large study, most students act independently of the others.
- ✓ **Expected Cell Frequency Condition:** All ≥ 5
- Use the χ^2 -model with $df = 6$ and do a chi-square test for homogeneity.



Students' Choices Same Across Colleges?

- **Mechanics:** I used the computer to produce the expected counts, χ^2 , and P-value.



- $\chi^2 \approx 93.66$
- P-value < 0.0001

Contingency table results:

Rows: Plan

Columns: None

Cell format

Count

Expected count

	Ag	ArtSci	Engineering	ILR	Total
Employed	209 210.8	198 227.7	177 176	101 70.57	685
Grad School	104 143.4	171 154.9	158 119.7	33 48.01	466
Other	135 93.85	115 101.4	39 78.34	16 31.42	305
Total	448	484	374	150	1456

Chi-Square test:

Statistic	DF	Value	P-value
Chi-square	6	93.65667	<0.0001

Students' Choices Same Across Colleges?

- **Conclusion:** The P-value is very small, so I reject the null hypothesis.
- I conclude that there's evidence that the post-graduation activities of students from these four colleges don't have the same distribution.

24.3

Examining the Residuals

Examining the Residuals

- After rejecting H_0 , we wonder where things differed.
- The residuals can be helpful.
- Even better:

standardized residuals: $c = \frac{Obs - Exp}{\sqrt{Exp}}$

- Standardized residuals are like z-scores.
- Tells us how far the observed is above or below the expected.

Standardized Residuals for College Grads

	Ag	A&S	Eng	ILR
Employed	−0.121866	−1.96860	0.078805	3.62235
Grad School	−3.28909	1.29304	3.50062	−2.16607
Other	4.24817	1.35192	−4.44511	−2.75117

- Engineering students: c is big, far from the expected value
- → They disproportionately go to grad school and not pursue other activities.
- Agriculture students disproportionately pursue other activities.

24.4

Chi-Square Test of Independence

Hepatitis and Tattoos

The contingency table shows the results of a study that looked at where people had their tattoo and whether they have Hepatitis C.

- Two categorical variables

	Hepatitis C	No Hepatitis C	Total
Tattoo, Parlor	17	35	52
Tattoo, Elsewhere	8	53	61
None	22	491	513
Total	47	579	626

- Contingency tables are used to see if one categorical variable is contingent on another.
- Are *Hepatitis C* and *Tattoo Status* independent?

Independence

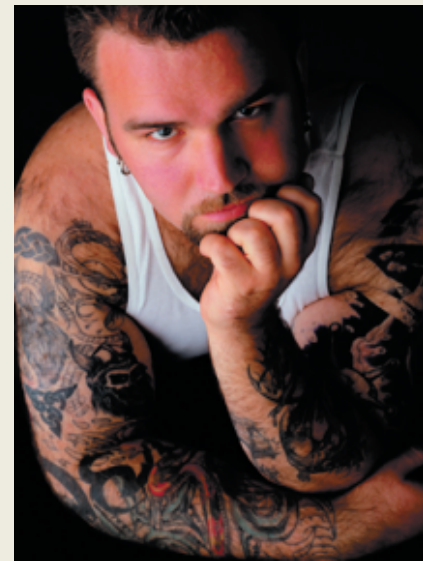
- Two events are **independent** if $P(A|B) = P(A)$.
- If *Tattoo Status* and *Hepatitis C* are **independent**, we would expect the proportions of “positive” to be the same for the three levels of *Tattoo Status*.
- Same criterion as homogeneity
- The nuances of homogeneity and independence statements are different.
- Same distribution vs. independent

Assumptions and Conditions

- **Counted Data Condition:** Always use counts, not proportions, percents, or measurements.
- **Independence Assumption:** The data were collected independently.
- **Expected Cell Frequency Condition:** All expected counts at least 5.
- **10% Condition:** Data come from a random sample of less than 10% of the population.

Tattoos and Hepatitis

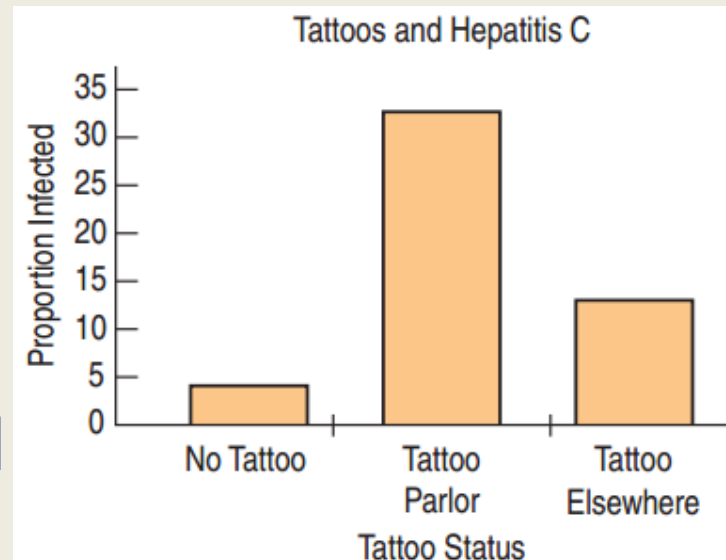
Are *Tattoo Status* and *Hepatitis Status* independent?



- Plan: Test for independence. I have a contingency table of 626 patients.
- Hypotheses:
 - H_0 : *Tattoo Status* and *Hepatitis Status* are independent.
 - H_A : *Tattoo Status* and *Hepatitis Status* are not independent.

Tattoos and Hepatitis

- **Model:** The bar chart suggests a difference in Hepatitis C based on tattoo status.



- ✓ **Counted Data Condition:** I have counts of individuals categorized on two variables.
- ✓ **Independence Assumption:** Not SRS, but likely independent.

Tattoos and Hepatitis

- **Model: (Continued)**

	Hepatitis C	No Hepatitis C	Total
Tattoo, Parlor	17 3.904	35 48.096	52
Tattoo, Elsewhere	8 4.580	53 56.420	61
None	22 38.516	491 474.484	513
Total	47	579	626

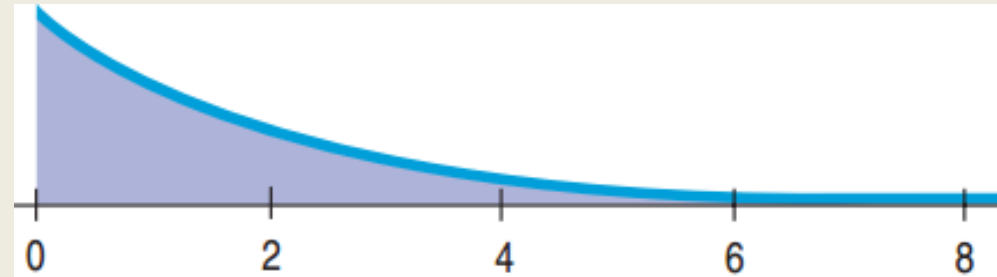
- × **Expected Cell Frequency Condition:**

Two expected counts < 5.

- Not all the assumptions are met, but I will carefully go ahead with the chi-square test for independence with $df = (3 - 1) \times (2 - 1) = 2$. I will check the residuals carefully.

Tattoos and Hepatitis C

- Mechanics: With df = 2 the χ^2 graph looks very different.
- Use computer just like with Homogeneity.



Chi-Square test:

Statistic	DF	Value	P-value
Chi-square	2	57.912174	<0.0001

- $\chi^2 = 57.91$, P-value < 0.0001

Tattoos and Hepatitis C

- **Conclusion:** With P-value < 0.0001 , reject H_0 . *Tattoo Status* and *Hepatitis Status* are not independent.
- Since the expected counts are < 5 , I need to check that the cells with small expected counts did not influence the results too greatly.

Examining the Residuals

	Hepatitis C	No Hepatitis C
Tattoo, Parlor	6.628	−1.888
Tattoo, Elsewhere	1.598	−0.455
None	−2.661	0.758

- Parlor/Hepatitis C residual is large positive. More Hepatitis C from those with tattoos from parlors than expected.
- None/Hepatitis C residual large negative. Less Hepatitis C from those without tattoos than expected.
- Cell counts small for Parlor/Hepatitis C. We should report this as a warning, or rethink the data.
- How can we increase cell counts?

Combining Groups

	Hepatitis C	No Hepatitis C	Total
Tattoo	25	88	113
None	22	491	513
Total	47	579	626

- Combining all those with tattoos gives large enough counts.
- New $\chi^2 = 42.42$ (df = 1)
- New P-value < 0.0001
- We conclude that *Tattoo Status* and *Hepatitis C Status* are not independent.
- We are concerned, but need more evidence to suggest that tattoo parlors are a problem.

Race and Traffic Stops

- Data were collected on drivers' race and whether they were searched when an officer stopped them.

- What test to decide if race is a factor in vehicle searches?

		Race			Total
		Black	White	Other	
Search	No	787	594	27	1408
	Yes	813	293	19	1125
	Total	1600	887	46	2533

- Two categorical variables: *Race* and *Search*.
- Chi-Square test for independence.
 - H_0 : *Race* and *Search* are independent.
 - H_A : *Race* and *Search* are not independent.

Chi-Square Mechanics

		Race			Total
		Black	White	Other	
Search	No	787	594	27	1408
	Yes	813	293	19	1125
	Total	1600	887	46	2533

Find the df, expected frequency for Black drivers who were stopped, that cell's component of χ^2 , and its residual.

- $df = (2 - 1)(3 - 1) = 2$
- Expected Frequency: $\frac{1125}{2533} \times 1600 \approx 710.62$
- χ^2 Contribution: $\frac{(813 - 710.62)^2}{710.62} \approx 14.75$
- S.Residual: $\frac{(813 - 710.62)}{\sqrt{710.62}} \approx 3.84$

Race and Traffic Stop Searches

- P-value < 0.0001
- There is strong evidence to conclude that police decisions to search are associated with race.
- Large residuals for Whites suggests that police search Whites less than independence would predict.
- It appears that Black drivers' cars are searched more often.

		Race		
		Black	White	Other
Search	No	-3.43	4.55	0.28
	Yes	3.84	-5.09	-0.31



Chi-Square and Causation

- Just like correlation, rejecting H_0 for a chi-square test does not imply causation.
- Does having Hepatitis C cause a craving to get a tattoo?
- Maybe there is a lurking variable.
- Is there a subculture that tends to have both?
- Just state independence, not causation.

What Can Go Wrong?

Don't use chi-square methods unless you have counts.

- Convert percents and proportions back to counts.

Beware large samples.

- There are no confidence intervals to see “how not independent” the categories are. Rarely are categories perfectly independent.

Don't say that one variable “depends” on the other just because they're not independent

- “Depends” suggest causation. Just state that the categories are “not independent” or are “associated.”