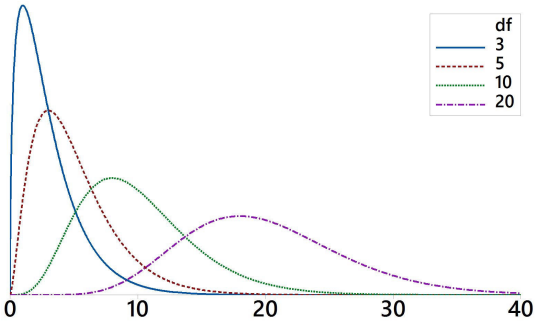
Chi-Squared Tests

- Chi-Squared Distribution
- Goodness of Fit Test
- Standardized Residuals
- Test for Independence

University of Pittsburgh

Chi-Squared Distribution

- **Chi-squared distribution:** continuous probability distribution with the following properties:
  - Unimodal, right-skewed, and always non-negative
  - Values get larger as degrees of freedom increase
  - Denoted by χ^2_v where v is the number of degrees of freedom
  - Degrees of freedom dependent upon number of categories in variable(s)

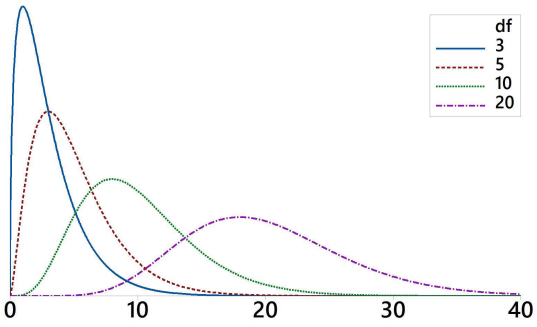


Note: Like the t-distribution, software must be used to calculate p-values because the shape of the distribution depends on the degrees of freedom.

Chi-Squared Distribution Table

Right Tail Probability	0.10	0.05	0.025	0.01	0.005
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

Table Continues



- Because the chi-squared distribution is not symmetric, the table only gives statistics for areas in the **upper** tail.
  - **Note:** All chi-squared tests in this course are upper one-sided.
  - As df increase, larger chi-square values are needed to get the same area in the upper tail.
  - For chi-squared tests, large test statistics are evidence against the null hypothesis.

Motivation: Goodness of Fit Test

- **Scenario:** MLB pitcher Mitch Keller throws four pitches: fastball, slider, curveball, and changeup. However, we are not interested in if they are all thrown with the same proportion.

Pitch	Fastball	Slider	Curveball	Changeup
Actual Count	114	44	30	12

- **Question:** Is there a 4:3:2:1 ratio between Keller's use of the fastball, slider, curveball, and changeup respectively?
- **Problem:** Variable is _____, but has _____ categories
  - One-sample proportion test _____
- **Solution:** Compare _____ counts against what we would expect under the _____

Goodness of Fit Test: Hypotheses and Conditions

Step	Description
Used for	Determining if a sample is consistent with a prespecified probability distribution
Hypotheses Framework	H_0: The specified probability distribution fits the data well H_A: The specified probability distribution does not fit the data well
Expected Count	np_i where p_i is the hypothesized proportion *Number of observations expected in each category under the assumption that the probability distribution is correct*
Conditions	Expected count for each category is at least 5

Example: Goodness of Fit Test

- **Question:** Is there a 4:3:2:1 ratio between Keller's use of the fastball, slider, curveball, and changeup using $\alpha = 0.05$?

Pitch	Fastball	Slider	Curveball	Changeup
Actual Count	114	44	30	12
Expected Count				

- **Hypotheses:**
  - H_0: $p_{FB} =$ _____, $p_{SL} =$ _____, $p_{CU} =$ _____, $p_{CH} =$ _____
  - H_A: At least _____ proportion differs from its _____
- **Conditions:** _____
  - All _____ counts are _____

Goodness of Fit Test: Test Statistic and P-Value

Step	Description
Test Statistic	$X^2 = \sum_{i=1}^{k} \frac{(\text{Observed} - \text{Expected})^2}{\text{Expected}}$ with $k - 1$ df
Idea	Compare observed and expected counts in each category Large differences between observed and expected counts ↪ Large chi-squared components ↪ Large test statistic Small p-values result from large test statistics ↪ Upper one-sided test
Conclusion	If H_0 is not rejected, then **all** of the hypothesized proportions are plausible If H_0 is rejected, then **at least one** of the hypothesized proportions is incorrect

Example: Goodness of Fit Test (Cont.)

- Summary:

Pitch	Fastball	Slider	Curveball	Changeup
Actual Count	114	44	30	12
Expected Count	80	60	40	20

- Test Statistic:

$X^2 =$ _____ + _____ + _____ + _____

$=$ _____

$=$ _____

- Degrees of Freedom: $k =$ _____ $=$ ___

Example: Goodness of Fit Test (Cont.)

- P-Value: $p = P($_____$) =$ _____
  - **Interpretation:** If Keller's pitches actually _____, then the probability we would get a _____ as extreme or more extreme than what we observed is _____.
  - **R-Function:** `1-pchisq(24.42,3)`

- **Conclusion:** _____ and conclude that the _____ _____ of at least one pitch is significantly different from the _____. (i.e. Keller's pitch distribution _____.)
  - $p =$ _____

Standardized Residuals

- **Standardized residual:** a measure of how many standard deviations an observed count is from its corresponding expected count for a particular category

$$d = \frac{\text{Observed} - \text{Expected}}{\sqrt{\text{Expected}}}$$

  - Standardized residuals:
    - Greater than 2 in magnitude are evidence that the observed and hypothesized proportions for that category are **significantly different**
    - Between 1 and 2 in magnitude are somewhat unusual but not sufficient evidence that the proportions are significantly different
    - Less than 1 in magnitude indicate that the observed and hypothesized proportions are quite close

Example: Standardized Residuals

- **Question:** Which pitches differ significantly from their hypothesized proportions?
- **Answer:** _____ (_____) and _____ (_____)

Pitch	Observed	Expected	Standardized Residual	Description
Fastball	114	80	$\frac{114-80}{\sqrt{80}} = \_\_\_\_\_$	
Slider	44	60	$\frac{44-60}{\sqrt{60}} = \_\_\_\_\_$	
Curveball	30	40	$\frac{30-40}{\sqrt{40}} = \_\_\_\_\_$	
Changeup	12	20	$\frac{12-20}{\sqrt{20}} = \_\_\_\_\_$	

Motivation: Test for Independence

- **Scenario:** Randomly sample 400 people and record their sex and handedness (left or right)

	Left-Handed	Right-Handed	Total
Male	20	140	160
Female	20	220	240
Total	40	360	400

- **Question:** Are sex and handedness independent or is one sex more likely to be left or right-handed?
- **Strategy:**
  - Assume sex and handedness are _____
  - Compare number of left and right-handed males and females against the number we would _____ in the sample if there was _____

Test for Independence

Step	Description
Used for	Determining if two categorical variables are independent
Hypotheses Framework	H_0: The variables are independent H_A: The variables are not independent
Expected Count	$e_{ij} = \frac{(\text{Row } i \text{ sum})(\text{Column } j \text{ sum})}{\text{Total sample size}}$ *Number of observations expected in each cell under the assumption that the variables are independent*
Conditions	Expected count for each category is at least 10
Test Statistic	$X^2 = \sum_{\text{All cells}} \frac{(\text{Observed} - \text{Expected})^2}{\text{Expected}}$ with $(r-1)(c-1)$ df r: Categories in row variable; c: Categories in column variable

Example: Test for Independence

- **Question:** Are sex and handedness independent?

Observed	Left	Right	Total
Male	20	140	160
Female	20	220	240
Total	40	360	400

Expected	Left	Right	Total
Male			160
Female			240
Total	40	360	400

- Hypotheses:
  - H_0: _____
  - H_A: _____

- Expected Count Condition: _____
  - All expected counts are _____

Example: Test for Independence

Observed	Left	Right	Total
Male	20	140	160
Female	20	220	240
Total	40	360	400

Expected	Left	Right	Total
Male	$\frac{(160)(40)}{400} = 16$	$\frac{(160)(360)}{400} = 144$	160
Female	$\frac{(240)(40)}{400} = 24$	$\frac{(240)(360)}{400} = 216$	240
Total	40	360	400

- Test Statistic:

$X^2 =$ _____ + _____ + _____ + _____

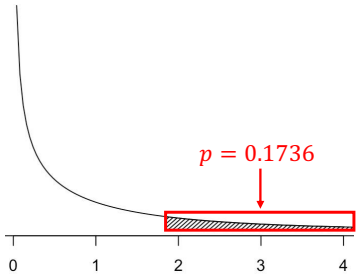
$=$ _____

$=$ _____

- Degrees of Freedom: $df =$ _____

EXAMPLE: TEST FOR INDEPENDENCE

- **P-Value:** _____
  - **R Function:** `1-pchisq(1.852,1)`



- **Conclusion:** _____ and conclude that there is _____ between sex and handedness.

- **Question:** Should we analyze the standardized residuals?
- **Answer:** _____
  - Conclusion of _____ implies that the observed and expected counts were _____
  - None of the standardized residuals would be _____

EXAMPLE: TEST FOR INDEPENDENCE USING R

- **Scenario:** Recall the example from Lecture 10 where a random sample of people were surveyed regarding their highest level of education and satisfaction with their job.

	Satisfied	Neutral	Not Satisfied	Total
High School	81	125	187	**393**
Bachelor's	42	55	57	**154**
Advanced	33	20	12	**65**
Total	**156**	**200**	**256**	**612**

- **Question:** Are education and job satisfaction related at the 5% level of significance?
- **Hypotheses:**
  - H_0: Education level and job satisfaction are _____
  - H_A: Education level and job satisfaction are _____

EXAMPLE: TEST FOR INDEPENDENCE USING R

- To obtain the observed and expected counts, read in the data and create the cross-classification table:

```
job_sat = read.table("job_sat.csv",
  header = TRUE, sep = ",")

js_table = table(job_sat)
```

	education	satisfaction
1	High School	Not Satisfied
2	High School	Neutral
3	Bachelors	Not Satisfied
4	High School	Not Satisfied
5	High School	Not Satisfied
6	Bachelors	Satisfied
7	High School	Not Satisfied
8	Advanced	Neutral
9	High School	Not Satisfied
10	Advanced	Neutral

- Use the **chisq.test** function on the table and **save** the results to a new variable.

```
results = chisq.test(js_table)
```

EXAMPLE: TEST FOR INDEPENDENCE USING R

- To obtain the test results, simply type the variable name:

```
> results

        Pearson's Chi-squared test

data:  js_table
X-squared = 33.205, df = 4, p-value = 1.084e-06
```

- To obtain the observed and expected counts, use the following:

`results$observed`

```
> results$observed
             satisfaction
education     Neutral Not Satisfied Satisfied
  Advanced          20           12        33
  Bachelors         55           57        42
  High School      125          187        81
```

`results$expected`

```
> results$expected
             satisfaction
education       Neutral Not Satisfied Satisfied
  Advanced      21.24183      27.18954  16.56863
  Bachelors     50.32680      64.41830  39.25490
  High School  128.43137     164.39216 100.17647
```


EXAMPLE: TEST FOR INDEPENDENCE USING R

- **Expected Count Condition:** _____
  - All expected counts are _____
- **Test Statistic:** _____
- **P-Value:** _____

```
> results$expected
             satisfaction
education       Neutral Not Satisfied Satisfied
  Advanced      21.24183      27.18954  16.56863
  Bachelors     50.32680      64.41830  39.25490
  High School  128.43137     164.39216 100.17647

        Pearson's Chi-squared test

data:  js_table
X-squared = 33.205, df = 4, p-value = 1.084e-06
```

- **Conclusion:** _____ and conclude that highest level of education and job satisfaction are _____.
  - To understand how the variables are related, analyze the _____ _____.

EXAMPLE: TEST FOR INDEPENDENCE USING R

- To determine how the variables are related, analyze the standardized residuals using the function `results$residuals`.

```
             satisfaction
education     Neutral Not Satisfied Satisfied
  Advanced        -0.269       -2.913     4.037
  Bachelors        0.659       -0.924     0.438
  High School     -0.303        1.763    -1.916
```

- **Question:** Which combinations had counts that were considerably higher than expected under the assumption of independence?
- **Answer:** Combinations with relatively _____ residuals
  - _____ degrees who were _____ → _____
  - _____ degrees who were _____ → _____

EXAMPLE: TEST FOR INDEPENDENCE USING R

- To determine how the variables are related, analyze the standardized residuals using the function `results$residuals`.

	satisfaction		
education	Neutral	Not Satisfied	Satisfied
Advanced	-0.269	-2.913	4.037
Bachelors	0.659	-0.924	0.438
High School	-0.303	1.763	-1.916

- **Question:** Which combinations had counts that were considerably lower than expected under the assumption of independence?
- **Answer:** Combinations with relatively _____ residuals
  - _____ degrees who were _____ → _____
  - _____ degrees who were _____ → _____

EXAMPLE: TEST FOR INDEPENDENCE USING R

- **Question:** How are education and job satisfaction related?
- **Answer:** _____ levels of education are associated with _____ job satisfaction
  - People with ____ education are more likely to be _____ with their job
  - Only a noticeable difference at the _____
    - "Neutral" job satisfaction is _____ across _____
    - For those with Bachelor's degrees, observed and expected counts are _____ _____ for _____

Observed

	satisfaction		
education	Neutral	Not Satisfied	Satisfied
Advanced	20	12	33
Bachelors	55	57	42
High School	125	187	81

Expected

	satisfaction		
education	Neutral	Not Satisfied	Satisfied
Advanced	21.24183	27.18954	16.56863
Bachelors	50.32680	64.41830	39.25490
High School	128.43137	164.39216	100.17647

Standardized Residuals

	satisfaction		
education	Neutral	Not Satisfied	Satisfied
Advanced	-0.269	-2.913	4.037
Bachelors	0.659	-0.924	0.438
High School	-0.303	1.763	-1.916