Math 3081 (Probability and Statistics)

Lecture #27 of 27 $\sim$ August 19th, 2021

The χ^2 Tests for Goodness of Fit and Independence

- The χ^2 Test for Goodness of Fit
- The χ^2 Test for Independence

This material represents §5.2.3-5.2.4 from the course notes, and problems 18-20 from WeBWorK 7.

Formerly Erstwhile Previously Earlier, I

Recall the χ^2 distribution:

Definition

The χ^2 distribution with k degrees of freedom is the continuous random variable Q_k whose probability density function $p_{Q_k}(x) = \dfrac{1}{2^{k/2}\Gamma(k/2)} \cdot x^{(k/2)-1}e^{-x/2}$ *for all real numbers* $x > 0$.

It is obtained by summing squares of independent standard normals:

Proposition (χ^2 Distribution From Normals)

If $X_1, \dots, X_n$ are independent standard normal random variables (i.e., with mean 0 and standard deviation 1), then the random variable $Q_n = X_1^2 + \cdots + X_n^2$ has a χ^2 distribution with n degrees of freedom.

Formerly Erstwhile Previously Earlier, II

Here is its main property as a sampling distribution:

Theorem (χ^2 Distribution As Sampling Distribution)

Suppose $n \geq 2$ and that $X_1, X_2, \ldots, X_n$ are independent, identically normally distributed random variables with mean μ and standard deviation σ. If $\overline{X} = \frac{1}{n}(X_1 + \cdots + X_n)$ denotes the sample mean and $S^2 = \dfrac{1}{n-1}\left[(X_1 - \overline{X})^2 + (X_2 - \overline{X})^2 + \cdots + (X_n - \overline{X})^2\right]$ denotes the sample variance, then the distribution of the test statistic $\dfrac{(n-1)S^2}{\sigma^2}$ is the χ^2 distribution Q_{n-1} with $n-1$ degrees of freedom.

Formerly Erstwhile Previously Earlier, III

The following theorem of Pearson gives a χ^2 test statistic for the scenario where values are drawn from a discrete random variable:

Theorem (χ^2 Goodness of Fit)

Suppose that a discrete random variable E has outcomes $e_1, e_2, \dots, e_k$ with respective probabilities $p_1, p_2, \dots, p_k$. If we sample this random variable n times, obtaining the respective outcomes $e_1, e_2, \dots, e_k$ a total of $x_1, x_2, \dots, x_k$ times, then as $n \to \infty$ the random variable $D = \dfrac{(x_1 - np_1)^2}{np_1} + \dfrac{(x_2 - np_2)^2}{np_2} + \cdots + \dfrac{(x_k - np_k)^2}{np_k}$ is χ^2-distributed with $k - 1$ degrees of freedom.

Formerly Erstwhile Previously Earlier, IV

Using this theorem, we can give a hypothesis testing procedure for analyzing the goodness of fit of a model:

- We take our test statistic as

$$\begin{aligned} d &= \frac{(x_1 - np_1)^2}{np_1} + \frac{(x_2 - np_2)^2}{np_2} + \cdots + \frac{(x_k - np_k)^2}{np_k} \\ &= \sum_{\text{data}} \frac{[\text{Observed} - \text{Expected}]^2}{\text{Expected}}. \end{aligned}$$

- Our hypotheses are usually $H_0 : d = 0$ and $H_a : d > 0$, since the value $d = 0$ means the model is perfect and a positive value of d indicates deviation from the model.

Formerly Erstwhile Previously Earlier, V

In order to apply Pearson's result above, we must verify that most of the predicted observation sizes np_i are at least 5.

- We will adopt the convention that at least 80% of the entries should be at least 5 or larger. (Another option is to combine some of these small entries into groups that have a predicted size greater than 5.)
- If the hypotheses are satisfied, then the test statistic is χ^2-distributed with $k - 1$ degrees of freedom, and we can calculate the p-value as $P(Q_{k-1} \geq d)$.

We then compare the p-value to the significance level and then either reject or fail to reject the null hypothesis, as usual.

Formerly Erstwhile Previously Earlier, VI

Our test is set up so that data perfectly fitting the model are not rejected, only data that are far away from the prediction.

- However, in some situations, we may instead want to test whether a model is “too good to believe” (e.g., if we are investigating whether it is reasonable to think that the data have been falsified or altered to adhere too closely to a model).
- In those situations we would instead want the hypotheses to be $H_0 : d = c$ and $H_a : d < c$ for (an arbitrary) positive c, and we would compute the p-value instead as $P(Q_{k-1} \leq d)$.

More χ^2 for Goodness of Fit, I

Example: To determine whether a pollster is actually conducting their polls, the tenths-place digits from a random sample of 200 of their reported results are tabulated. The results are given below. It is expected that the tenths-place digit from poll percentages of thousands of people should be essentially uniformly distributed. Test at the 10%, 1%, and 0.02% significance levels whether the data appear to adhere to a uniform model.

Tenths Digit	0	1	2	3	4	5	6	7	8	9
Observed	7	26	13	44	25	10	9	41	12	13

More χ^2 for Goodness of Fit, I

Example: To determine whether a pollster is actually conducting their polls, the tenths-place digits from a random sample of 200 of their reported results are tabulated. The results are given below. It is expected that the tenths-place digit from poll percentages of thousands of people should be essentially uniformly distributed. Test at the 10%, 1%, and 0.02% significance levels whether the data appear to adhere to a uniform model.

Tenths Digit	0	1	2	3	4	5	6	7	8	9
Observed	7	26	13	44	25	10	9	41	12	13

- We will add the "Expected" and $(O - E)^2/E$ rows to the table, and then perform the hypothesis test.

More χ^2 for Goodness of Fit, II

Example: To determine whether a pollster is actually conducting their polls, the tenths-place digits from a random sample of 200 of their reported results are tabulated. Test at the 10%, 1%, and 0.02% significance levels whether the data appear to adhere to a uniform model.

Digit	0	1	2	3	4	5	6	7	8	9
Observed	7	26	13	44	25	10	9	41	12	13
Expected	20	20	20	20	20	20	20	20	20	20
$(O-E)^2/E$	8.35	1.8	2.45	28.8	1.25	5	6.05	22.05	3.2	2.45

More χ^2 for Goodness of Fit, II

Example: To determine whether a pollster is actually conducting their polls, the tenths-place digits from a random sample of 200 of their reported results are tabulated. Test at the 10%, 1%, and 0.02% significance levels whether the data appear to adhere to a uniform model.

Digit	0	1	2	3	4	5	6	7	8	9
Observed	7	26	13	44	25	10	9	41	12	13
Expected	20	20	20	20	20	20	20	20	20	20
$(O-E)^2/E$	8.35	1.8	2.45	28.8	1.25	5	6.05	22.05	3.2	2.45

- Our test statistic is $d = 8.45 + 1.8 + 2.45 + \cdots + 2.45 = 81.5$.
- There are 10 possible outcomes hence $10 - 1 = 9$ degrees of freedom.
- Thus, the p-value is $P(Q_9 \geq 81.5) = 8.13 \cdot 10^{-14}$. This is extremely small, so we reject the null hypothesis at all of the indicated significance levels.

More χ^2 for Goodness of Fit, III

Example: The pollster, in response to the accusations from the previous test, defends their innocence by sending another sample of 200 polls. Test at the 10%, 1%, and 0.02% significance levels whether the results are believable.

Digit	0	1	2	3	4	5	6	7	8	9
Observed	22	21	20	20	19	18	20	19	21	20

More χ^2 for Goodness of Fit, III

Example: The pollster, in response to the accusations from the previous test, defends their innocence by sending another sample of 200 polls. Test at the 10%, 1%, and 0.02% significance levels whether the results are believable.

Digit	0	1	2	3	4	5	6	7	8	9
Observed	22	21	20	20	19	18	20	19	21	20

- We would expect (with a perfect model) to get an entry of 20 in each row.
- This time, the numbers are all very suspiciously close to 20. We will now test whether the uniform model is too accurate (with the left tail for the distribution rather than the right tail).

More χ^2 for Goodness of Fit, IV

Example: The pollster, in response to the accusations from the previous test, defends their innocence by sending another sample of 200 polls. Test at the 10%, 1%, and 0.02% significance levels whether the results are believable.

Digit	0	1	2	3	4	5	6	7	8	9
Observed	22	21	20	20	19	18	20	19	21	20
Expected	20	20	20	20	20	20	20	20	20	20
$(O-E)^2/E$	0.2	0.05	0	0	0.05	0.2	0	0.05	0.05	0

More χ^2 for Goodness of Fit, IV

Example: The pollster, in response to the accusations from the previous test, defends their innocence by sending another sample of 200 polls. Test at the 10%, 1%, and 0.02% significance levels whether the results are believable.

Digit	0	1	2	3	4	5	6	7	8	9
Observed	22	21	20	20	19	18	20	19	21	20
Expected	20	20	20	20	20	20	20	20	20	20
$(O-E)^2/E$	0.2	0.05	0	0	0.05	0.2	0	0.05	0.05	0

- Our test statistic is $d = 0.2 + 0.05 + 0 + \cdots + 0 = 0.6$.
- As before there are $10 - 1 = 9$ degrees of freedom.
- Thus, the p-value is $P(Q_9 \leq 0.6) = 6.64 \cdot 10^{-5}$. This is once again extremely small, so we reject the null hypothesis at all of the indicated significance levels.

More χ^2 for Goodness of Fit, V

The last two examples demonstrate how we can use the χ^2 test to assess whether data have been obviously forged: we can check whether they deviate too much from an expected model, and also whether they are too close to an expected model.

- Of course, in the latter case, it is important to be careful about interpreting the p-value appropriately.
- Large samples may tend to make the p-value smaller (indicating close adherence to the model).
- To calibrate one's sense of the p-value here, one may do simulations with randomly-generated data of the same sample size, to look at the actual distributions of p-values.

More χ^2 for Goodness of Fit, VI

Since we're discussing digits, I'll also mention another distribution of digits one should tend to see in certain situations.

- For quantities that range over several orders of magnitude, their leading digits should not be uniformly distributed, but rather follow Benford's law:

Digit	1	2	3	4	5	6	7	8	9
Benford	30.1%	17.6%	12.5%	9.7%	7.9%	6.7%	5.8%	5.1%	4.6%

- The idea is that the base-10 logarithm of the number (modulo 1) should be approximately uniform, meaning that the leading digit d should appear with probability $\log_{10}(1 + 1/d)$.
- Benford's law applies to quantities such as stock prices, house prices, population numbers, and lengths of rivers.
- One can prove that leading digits of fast-growing sequences like the factorials and powers of 2 obey Benford's law.

More χ^2 for Goodness of Fit, VII

Example: It is believed that a Poisson model is appropriate to model the number of collisions at a particular busy intersection in a given week. The collisions are tabulated over a 5-year period (a total of 261 weeks), and the results are given below. Test at the 9% and 1% significance levels the accuracy of the model

1. with parameter $\lambda = 2.2$.
2. with parameter $\lambda = 2.9$.

# Collisions	0	1	2	3	4	5	6	7+
Observed	17	45	66	55	38	21	12	7

More χ^2 for Goodness of Fit, VII

Example: It is believed that a Poisson model is appropriate to model the number of collisions at a particular busy intersection in a given week. The collisions are tabulated over a 5-year period (a total of 261 weeks), and the results are given below. Test at the 9% and 1% significance levels the accuracy of the model

1. with parameter $\lambda = 2.2$.
2. with parameter $\lambda = 2.9$.

# Collisions	0	1	2	3	4	5	6	7+
Observed	17	45	66	55	38	21	12	7

- If the Poisson model is accurate, we would expect the proportion of outcomes yielding d collisions to be $\dfrac{\lambda^d e^{-\lambda}}{d!}$, so the expected number of occurrences would be 261 times this quantity.

More χ^2 for Goodness of Fit, VIII

Example: Collisions are tabulated over a 5-year period (a total of 261 weeks), and the results are given below. Test at the 9% and 1% significance levels the accuracy of a Poisson model

1. with parameter $\lambda = 2.2$.

Collisions	0	1	2	3	4	5	6	7+
Observed	17	45	66	55	38	21	12	7
Expected	28.92	63.63	69.99	51.32	28.23	12.42	4.55	1.95
$(O-E)^2/E$	4.9128	13.7927	0.2270	0.2635	3.3833	5.9271	12.174	13.109

More χ^2 for Goodness of Fit, VIII

Example: Collisions are tabulated over a 5-year period (a total of 261 weeks), and the results are given below. Test at the 9% and 1% significance levels the accuracy of a Poisson model

1. with parameter $\lambda = 2.2$.

Collisions	0	1	2	3	4	5	6	7+
Observed	17	45	66	55	38	21	12	7
Expected	28.92	63.63	69.99	51.32	28.23	12.42	4.55	1.95
$(O - E)^2/E$	4.9128	13.7927	0.2270	0.2635	3.3833	5.9271	12.174	13.109

- Here, we have 2 entries out of 8 that are less than 5. This is a sufficiently large percentage that we can use our χ^2 test.
- Another option would be to combine the last two cells to make "6+" and then do the calculation.

More χ^2 for Goodness of Fit, IX

Example: Collisions are tabulated over a 5-year period (a total of 261 weeks), and the results are given below. Test at the 9% and 1% significance levels the accuracy of a Poisson model

1. with parameter $\lambda = 2.2$.

Collisions	0	1	2	3	4	5	6	7+
Observed	17	45	66	55	38	21	12	7
Expected	28.92	63.63	69.99	51.32	28.23	12.42	4.55	1.95
$(O-E)^2/E$	4.9128	13.7927	0.2270	0.2635	3.3833	5.9271	12.174	13.109

More χ^2 for Goodness of Fit, IX

Example: Collisions are tabulated over a 5-year period (a total of 261 weeks), and the results are given below. Test at the 9% and 1% significance levels the accuracy of a Poisson model

1. with parameter $\lambda = 2.2$.

Collisions	0	1	2	3	4	5	6	7+
Observed	17	45	66	55	38	21	12	7
Expected	28.92	63.63	69.99	51.32	28.23	12.42	4.55	1.95
$(O-E)^2/E$	4.9128	13.7927	0.2270	0.2635	3.3833	5.9271	12.174	13.109

- Our test statistic is $d = 4.9128 + 13.7927 + 0.2270 + \cdots + 13.1090 = 53.7898$.
- There are 8 outcomes hence $8 - 1 = 7$ degrees of freedom.
- Thus, the p-value is $P(Q_7 \geq 53.7898) = 2.588 \cdot 10^{-9}$.
- Since this is far below our significance levels, we reject the null hypothesis in both cases: the model appears incorrect.

More χ^2 for Goodness of Fit, X

Example: Collisions are tabulated over a 5-year period (a total of 261 weeks), and the results are given below. Test at the 9% and 1% significance levels the accuracy of a Poisson model

2. with parameter $\lambda = 2.9$.

Collisions	0	1	2	3	4	5	6	7+
Observed	17	45	66	55	38	21	12	7
Expected	14.36	41.65	60.39	58.38	42.32	24.55	11.86	7.50
$(O-E)^2/E$	0.4849	0.2699	0.5215	0.1952	0.4414	0.5125	0.0016	0.0327

More χ^2 for Goodness of Fit, X

Example: Collisions are tabulated over a 5-year period (a total of 261 weeks), and the results are given below. Test at the 9% and 1% significance levels the accuracy of a Poisson model

2. with parameter $\lambda = 2.9$.

Collisions	0	1	2	3	4	5	6	7+
Observed	17	45	66	55	38	21	12	7
Expected	14.36	41.65	60.39	58.38	42.32	24.55	11.86	7.50
$(O-E)^2/E$	0.4849	0.2699	0.5215	0.1952	0.4414	0.5125	0.0016	0.0327

- Our test statistic is $d = 0.4849 + \cdots + 0.0327 = 2.4597$.
- As above there are 7 degrees of freedom, so the p-value is $P(Q_7 \geq 2.4597) = 0.9301$.
- The p-value is now quite large, so we fail to reject the null hypothesis. The model seems quite good!

More χ^2 for Goodness of Fit, XI

In this last example, we could have performed a maximum likelihood estimation for the Poisson parameter to find the ideal λ fitting the observed data.

- The maximum likelihood estimator for that example ends up being $\hat{\lambda} = 2.7586$, which is not far from the actual value.
- However, if we do this sort of “tuning” of the model to fit the data, we would expect to get somewhat better agreement than without being able to adjust a parameter to get a better fit.

More χ^2 for Goodness of Fit, XII

To obtain reliable results, we must correct the χ^2 test in the situation where we select parameters to fit the data.

- The usual method of correction is as follows: if we use a model with r unknown parameters that have been calculated to obtain optimal fit to the observed data, we should use a χ^2 test with $k - 1 - r$ degrees of freedom (rather than $k - 1$).
- Roughly speaking, each unknown parameter removes one degree of freedom from the hypothesis test, since each parameter value we are allowed to choose will allow us to model one additional outcome from the list of k correctly.

χ^2 for Independence, I

As a final application of the χ^2 test, we will apply it to study the independence of discrete random variables.

- I wanted to do this as the last topic because I think it is a nice way of tying our course together, since it is about a topic from the very beginning of the term.
- Recall that we can test whether two discrete random variables X and Y are independent by checking whether the joint pdf $p_{X,Y}(x,y) = p_X(x) \cdot p_Y(y)$ is the product of the individual pdfs.
- If we construct a joint probability distribution table, we can check whether X and Y are independent by computing the row and column sums, and then testing whether each entry $p_{X,Y}(x,y)$ in the table is the product of its associated row sum $p_X(x)$ and its associated column sum $p_Y(y)$.

χ^2 for Independence, I

So suppose, now, that we are computing the joint distribution table for two random variables X and Y by sampling a population.

- We would expect the entries in the resulting table (which are now counts of individual observations) to show some random variation in their values away from the true proportion $p_{X,Y}(x,y)$.
- Thus, if we try to determine whether X and Y are independent using the criterion $p_{X,Y}(x,y) = p_X(x) \cdot p_Y(y)$, it is very unlikely that we would see exact independence.
- We can, however, adapt Pearson's χ^2 test for goodness-of-fit to give a hypothesis test for independence: the scenario we are describing is essentially identical to the one we just analyzed.

χ^2 for Independence, II

Here is the main theorem:

Theorem (χ^2 Independence)

Suppose that the discrete random variables X and Y have outcomes $x_1, \dots, x_a$ and $y_1, \dots, y_b$. Suppose that (X, Y) is sampled n times, such that the outcome x_i occurs a proportion p_i times, the outcome y_j occurs a proportion q_j times, and the outcome pair (x_i, y_j) occurs $a_{i,j}$ times for each $1 \leq i \leq a$ and $1 \leq j \leq b$. Then, as $n \to \infty$, the random variable $D = \displaystyle\sum_{i=1}^{a} \sum_{j=1}^{b} \frac{(a_{i,j} - np_iq_j)^2}{np_iq_j}$ is χ^2-distributed with $(a-1)(b-1)$ degrees of freedom.

χ^2 for Independence, III

Here is the rough idea:

- If X and Y are independent, then np_iq_j is the expected number of times we should obtain the outcomes x_i (probability p_i) and y_j (probability q_j) together.
- Thus, we are computing the same sum $D = \sum_{\text{data}} \frac{[\text{Observed} - \text{Expected}]^2}{\text{Expected}}$ as before.
- The proof of this result is similar to the one we gave earlier for goodness-of-fit: for large n, each of the ratios $\frac{(a_{i,j} - np_iq_j)^2}{np_iq_j}$ will behave like a scaled χ^2 distribution with 1 degree of freedom.

χ^2 for Independence, IV

Let me briefly try to explain the non-obvious fact about why the number of degrees of freedom is $(a-1)(b-1)$.

- Essentially, the idea is that if we are filling entries into the joint pdf table of X and Y, then all of the entries in the $a \times b$ table are completely determined once we fill in the upper left $(a-1) \times (b-1)$ table, under the presumption that we also know the row and column sums p_i and q_j (because we extract p_i and q_j from the data, we view them as parameters that we have selected).
- We can fill in all the entries because once we have all but one entry in a given row, we can fill in the last entry since we know the row sum. The same holds true for the columns, so applying this for each row and column (including the bottom row that we just filled) allows us to fill the entire grid.

χ^2 for Independence, V

- On the other hand, if we have fewer than $(a-1)(b-1)$ entries, we cannot fill the entire grid.
- Thus, the total number of independent values is $(a-1)(b-1)$, so this is the number of degrees of freedom.
- An equivalent (and more highbrow) way to make this observation is that the entries in the upper $(a-1) \times (b-1)$ subgrid form a basis for the vector space consisting of the entries of the grid with fixed row and column sums.

χ^2 for Independence, VI

Using the theorem, we can give a hypothesis testing procedure for analyzing the independence of two random variables X and Y:

- First, we write down the $a \times b$ joint probability distribution table for the observed values of X and Y, and compute the row proportions p_i and column proportions q_j.
- Then we compute the expected value of each entry np_iq_j, and calculate the test statistic as
$$d = \sum_{i=1}^{a}\sum_{j=1}^{b}\frac{(a_{i,j} - np_iq_j)^2}{np_iq_j} = \sum_{\text{data}}\frac{[\text{Observed} - \text{Expected}]^2}{\text{Expected}}.$$
- We take as our hypotheses $H_0 : d = 0$ and $H_a : d > 0$, since the value $d = 0$ means that the model is perfect (indicating that all of the entries are exactly equal to the predicted value, which means X and Y are independent) and a positive value of d indicates deviation from independence.

χ^2 for Independence, VII

In order to apply Pearson's theorem, we must verify that most of the predicted observation sizes np_i are at least 5.

- We will adopt the same convention as before that at least 80% of the entries should be at least 5 or larger.
- If that is the case, then the test statistic is χ^2-distributed with $(a-1)(b-1)$ degrees of freedom, and we can calculate the p-value as $P(Q_{(a-1)(b-1)} \geq d)$.

We then compare the p-value to the significance level and then either reject or fail to reject the null hypothesis, as usual.

χ^2 for Independence, VIII

Example: The faculty members in a university mathematics department are either tenure-track or non-tenure-track. These categories are broken down further by gender as indicated below. Test at the 9% and 0.8% significance levels whether the two variables of tenure track status and gender are independent.

Observed	Tenure-Track	Non-Tenure-Track
Male	20	8
Female	4	8

χ^2 for Independence, IX

Example: The faculty members in a university mathematics department are either tenure-track or non-tenure-track. These categories are broken down further by gender as indicated below. Test at the 9% and 0.8% significance levels whether the two variables of tenure track status and gender are independent.

Observed	Tenure-Track	Non-Tenure-Track
Male	20	8
Female	4	8

χ^2 for Independence, IX

Example: The faculty members in a university mathematics department are either tenure-track or non-tenure-track. These categories are broken down further by gender as indicated below. Test at the 9% and 0.8% significance levels whether the two variables of tenure track status and gender are independent.

Observed	Tenure-Track	Non-Tenure-Track
Male	20	8
Female	4	8

- There are 40 faculty in total, so we can compute the row and column proportions and then fill in the table of expected values

χ^2 for Independence, X

Example: The faculty members in a university mathematics department are either tenure-track or non-tenure-track. These categories are broken down further by gender as indicated below. Test at the 9% and 0.8% significance levels whether the two variables of tenure track status and gender are independent.

Observed	Tenure-Track	Non-Tenure-Track
Male	20	8
Female	4	8

We get the following table of expected results:

Expected	Tenure-Track	Non-Tenure-Track	Proportion
Male	$40 \cdot 0.42 = 16.8$	$40 \cdot 0.28 = 11.2$	0.7
Female	$40 \cdot 0.18 = 7.2$	$40 \cdot 0.12 = 4.8$	0.3
Proportion	0.6	0.4	

χ^2 for Independence, XI

Example: The faculty members in a university mathematics department are either tenure-track or non-tenure-track. These categories are broken down further by gender as indicated below. Test at the 9% and 0.8% significance levels whether the two variables of tenure track status and gender are independent.

Exp (Obs)	Tenure-Track	Non-Tenure-Track
Male	16.8 (20)	11.2 (8)
Female	7.2 (4)	4.8 (8)

χ^2 for Independence, XI

Example: The faculty members in a university mathematics department are either tenure-track or non-tenure-track. These categories are broken down further by gender as indicated below. Test at the 9% and 0.8% significance levels whether the two variables of tenure track status and gender are independent.

Exp (Obs)	Tenure-Track	Non-Tenure-Track
Male	16.8 (20)	11.2 (8)
Female	7.2 (4)	4.8 (8)

- The test statistic is
$$\frac{(20-16.8)^2}{16.8}+\frac{(8-11.2)^2}{11.2}+\frac{(4-7.2)^2}{7.2}+\frac{(8-4.8)^2}{4.8}=5.0794.$$
- The total number of degrees of freedom is $(2-1)(2-1)=1$.

χ^2 for Independence, XII

Example: The faculty members in a university mathematics department are either tenure-track or non-tenure-track. These categories are broken down further by gender as indicated below. Test at the 9% and 0.8% significance levels whether the two variables of tenure track status and gender are independent.

Exp (Obs)	Tenure-Track	Non-Tenure-Track
Male	16.8 (20)	11.2 (8)
Female	7.2 (4)	4.8 (8)

χ^2 for Independence, XII

Example: The faculty members in a university mathematics department are either tenure-track or non-tenure-track. These categories are broken down further by gender as indicated below. Test at the 9% and 0.8% significance levels whether the two variables of tenure track status and gender are independent.

Exp (Obs)	Tenure-Track	Non-Tenure-Track
Male	16.8 (20)	11.2 (8)
Female	7.2 (4)	4.8 (8)

- With $d = 5.0794$ and $df = 1$, the p-value is $P(Q_1 \geq 5.0794) = 0.02421$.
- Since the p-value is below the 9% significance level but above the 0.8% significance level, we reject the null hypothesis in the first case but not in the second case.
- Our interpretation of the test is that we have moderately strong evidence that the variables are not independent.

χ^2 for Independence, XIII

Example: A survey is taken of 400 households asking about the number of children and the number of TVs in the household. Test at the 11% and 2% significance levels whether the number of TVs is independent of the number of children.

Observed	0 Children	1 Child	2 Children	3+ Children
0 TVs	10	25	29	16
1 TV	19	88	104	29
2+ TVs	9	24	29	18

χ^2 for Independence, XIII

Example: A survey is taken of 400 households asking about the number of children and the number of TVs in the household. Test at the 11% and 2% significance levels whether the number of TVs is independent of the number of children.

Observed	0 Children	1 Child	2 Children	3+ Children
0 TVs	10	25	29	16
1 TV	19	88	104	29
2+ TVs	9	24	29	18

- As before, we compute the row and column proportions and then fill in the table of expected values.

χ^2 for Independence, XIV

Example: A survey is taken of 400 households asking about the number of children and the number of TVs in the household. Test at the 11% and 2% significance levels whether the number of TVs is independent of the number of children.

Observed	0 Children	1 Child	2 Children	3+ Children
0 TVs	10	25	29	16
1 TV	19	88	104	29
2+ TVs	9	24	29	18

Expected	0 Children	1 Child	2 Children	3+ Children	Proportion
0 TVs	7.6	27.4	32.4	12.6	0.2
1 TV	22.8	82.2	97.2	37.8	0.6
2+ TVs	7.6	27.4	32.4	12.6	0.2
Proportion	0.095	0.3425	0.405	0.1575	

χ^2 for Independence, XV

Example: Test at the 11% and 2% significance levels whether the number of TVs is independent of the number of children.

Exp (Obs)	0 Children	1 Child	2 Children	3+ Children
0 TVs	7.6 (10)	27.4 (25)	32.4 (29)	12.6 (16)
1 TV	22.8 (19)	82.2 (88)	97.2 (104)	37.8 (29)
2+ TVs	7.6 (9)	27.4 (24)	32.4 (29)	12.6 (18)

χ^2 for Independence, XV

Example: Test at the 11% and 2% significance levels whether the number of TVs is independent of the number of children.

Exp (Obs)	0 Children	1 Child	2 Children	3+ Children
0 TVs	7.6 (10)	27.4 (25)	32.4 (29)	12.6 (16)
1 TV	22.8 (19)	82.2 (88)	97.2 (104)	37.8 (29)
2+ TVs	7.6 (9)	27.4 (24)	32.4 (29)	12.6 (18)

- Then $d = \frac{(10-7.6)^2}{7.6} + \frac{(25-27.4)^2}{27.4} + \cdots + \frac{(18-12.6)^2}{12.6} = 9.1602$.
- The total number of degrees of freedom is $(4-1)(3-1) = 6$, so the p-value is given by $P(Q_6 \geq 9.1602) = 0.1648$.
- Since the p-value is above the 11% and 2% significance levels, we fail reject the null hypothesis in both cases
- Our interpretation is that we have fairly weak evidence that the variables are not independent. (It's not zero, but it's not very compelling.)

χ^2 for Independence, XVI

Example: A poll is taken on a trenchant political issue and the support is broken down by age group, as shown below. Test at the 8%, 2%, and 0.3% significance levels whether the level of support is independent of the age group.

Observed	Age 18-29	Age 30-49	Age 50-64	Age 65+
Support	20	13	12	8
Oppose	7	9	14	17

χ^2 for Independence, XVI

Example: A poll is taken on a trenchant political issue and the support is broken down by age group, as shown below. Test at the 8%, 2%, and 0.3% significance levels whether the level of support is independent of the age group.

Observed	Age 18-29	Age 30-49	Age 50-64	Age 65+
Support	20	13	12	8
Oppose	7	9	14	17

As before, we find the row and column proportions and then use them to fill in the table of expected values.

χ^2 for Independence, XVII

Example: A poll is taken on a trenchant political issue and the support is broken down by age group, as shown below. Test at the 8%, 2%, and 0.3% significance levels whether the level of support is independent of the age group.

Observed	Age 18-29	Age 30-49	Age 50-64	Age 65+
Support	20	13	12	8
Oppose	7	9	14	17

Expected	Age 18-29	Age 30-49	Age 50-64	Age 65+	Proportion
Support	14.31	11.66	13.78	13.25	0.53
Oppose	12.69	10.34	12.22	11.75	0.47
Proportion	0.27	0.22	0.26	0.25	

χ^2 for Independence, XVIII

Example: Test at the 8%, 2%, and 0.3% significance levels whether the level of support is independent of the age group.

Exp (Obs)	Age 18-29	Age 30-49	Age 50-64	Age 65+
Support	14.31 (20)	11.66 (13)	13.78 (12)	13.25 (8)
Oppose	12.69 (7)	10.34 (9)	12.22 (14)	11.75 (17)

χ^2 for Independence, XVIII

Example: Test at the 8%, 2%, and 0.3% significance levels whether the level of support is independent of the age group.

Exp (Obs)	Age 18-29	Age 30-49	Age 50-64	Age 65+
Support	14.31 (20)	11.66 (13)	13.78 (12)	13.25 (8)
Oppose	12.69 (7)	10.34 (9)	12.22 (14)	11.75 (17)

- The test statistic is $\frac{(20-14.31)^2}{14.31} + \cdots + \frac{(17-11.75)^2}{11.75} = 10.057$.
- The total number of degrees of freedom is $(4-1)(2-1) = 3$, so the p-value is given by $P(Q_3 \geq 10.057) = 0.01809$.
- Since the p-value is below the 8% and 2% significance levels, we reject the null hypothesis in those cases. It is above the 0.3% significance level, so we fail to reject there.
- Our interpretation of the test is that we have fairly strong evidence that the variables are not independent: the support does appear to depend on the age group.

Fisher's Exact Test, I

As a final example, I will discuss Fisher's exact test, which gives an exact hypothesis testing method for 2×2 tables without the need for a χ^2 approximation.

- In this case, there is only 1 degree of freedom.
- Per the discussion earlier, the idea is that if the row and column totals are known, then only the single upper-left entry is required to determine the full table.

Fisher's Exact Test, II

Fisher's original example was of the "lady tasting tea", who claimed to be able to decide, solely by the flavor, whether a cup of tea with milk had the milk poured into the tea or the tea poured into the milk.

- In the real experiment, eight cups of tea with milk were poured, four with milk first and four with tea first.
- The lady tasted each (under blind conditions) and decided whether the tea or the milk had been poured first.
- The problem is to decide, based on how many cups the lady identifies correctly, how plausible it is that she really can tell the difference.
- Under the null hypothesis of random guessing, we assume that the lady would guess exactly 4 cups of each type, since (as part of the test conditions) she is told that there will be 4 cups of each type.

Fisher's Exact Test, III

We can break down the results as follows:

Observed	Lady: Milk first	Lady: Tea first
Milk poured first	a	b
Tea poured first	c	d

- To obtain the table above the lady will always guess $a+c$ of the cups to have milk first and $b+d$ to have tea first, so there are a total $\binom{a+b+c+d}{a+c}$ possible tables satisfying this condition.
- To obtain the specific table above, exactly a of the $a+c$ cups the lady says have milk must actually have milk, and exactly d of the cups the lady says have tea must actually have tea.
- There are $\binom{a+c}{a} \cdot \binom{b+d}{d}$ ways of making these selections, so the total probability of obtaining the given table is $\binom{a+c}{a}\binom{b+d}{d}/\binom{a+b+c+d}{a+c}$.

Fisher's Exact Test, IV

We can then compute the probability of obtaining a result at least as close to completely accurate by summing over the possible tables with upper-left entry at least as large as the observed value.

- For example, if the results had been

Observed	Lady: Milk first	Lady: Tea first
Milk poured first	3	1
Tea poured first	1	3

  then the probability of obtaining this precise table is $\binom{4}{3}\binom{4}{3}/\binom{8}{4} = \frac{16}{70} \approx 0.2286$. The only result yielding more correct responses would be the table with entries $(4, 0), (0, 4)$ which occurs with probability $\binom{4}{4}\binom{4}{4}/\binom{8}{4} = \dfrac{1}{70} \approx 0.0143$.
- Thus, the tail probability is the sum $\frac{16}{70} + \frac{1}{70} \approx 0.2429$.
- We would not view this as conclusive evidence!

Fisher's Exact Test, V

In fact, the results of the actual test were that the lady correctly identified all 8 cups.

Observed	Lady: Milk first	Lady: Tea first
Milk poured first	4	0
Tea poured first	0	4

- In that case, the probability of obtaining the result by random guessing is $\binom{4}{4}\binom{4}{4}/\binom{8}{4} = \dfrac{1}{70} \approx 0.0143$.
- That is much more compelling evidence that she was not actually guessing.

We can use Fisher's exact test in scenarios with small sample sizes for 2×2 tables, and it often yields better results than the χ^2 approximation.

Closing Remarks, I

We are now at the end of our discussion of hypothesis tests.

- What I'd like to mention, though, is that there are lots of other hypothesis tests out there.
- Some of them, like the z test and t test, have broad applicability and are fairly robust. That's why we teach you about them in introductory courses, since you can get good mileage from them.
- Others, like Fisher's exact test, are for very specific situations. They are very good, but only for the particular situation in which they apply.
- It is very important to make sure that any hypotheses of the tests you do use are actually applicable! (For example, don't use a t test with exponentially-distributed data and a sample size $n = 5$.)

Closing Remarks, II

What I am hoping you'll take from all of this is the importance of asking the right question about the quantity you really want to understand, and using the most appropriate statistical methods available.

- Most modern statistical methods rely on computer simulation and numerical approximation to check whether the techniques are reliable.
- It is also entirely possible that you might end up wanting to ask something that isn't easily answered by any of the tests you've seen.
- You may end up having to search for an appropriate statistical method, or even try to design your own. (Not something for amateurs! Take more statistics classes if you're going to be doing serious statistical analysis in your research!)

Closing Remarks, III

You should also now feel comfortable reading actual uses of statistics in scientific papers.

- You may not be familiar with the precise tests being used, but you now should have a good grasp of the language and terminology.
- In particular, I would hope that you can now spot a lot of the misuses of statistics that I've spoken at length about.

One thing I'll call your attention to is $\sim$ very small sample sizes.

- Beware any results that have very small sample sizes, since (as we have seen) most tests have a very low power with a small sample: there is a high probability of getting unreliable results.
- Furthermore, populations tend to have more outliers than normal distributions would predict, and this will lower the accuracy of t tests.

Summary

We did more examples of the χ^2 test for goodness-of-fit.

We discussed the χ^2 test for independence.

We discussed Fisher's exact test.

Next lecture: There isn't one: the lectures are over! (Though the TA is holding a review session tomorrow and I am holding one on Saturday, both from 1-3pm.)

It’s The End!

We’re now at the end of the course (except of course for the last WeBWorK and the final).

I hope you enjoyed learning some probability and statistics with me this semester as much as I enjoyed teaching it. Although I am a number theorist and pure mathematician, I was also trained as a scientist and thus, teaching you about the probabilistic and statistical arts is something I enjoy getting to do.

In particular, I would hope that you now have a better understanding of the power, but also of the limitations, of hypothesis testing and of statistics, so that you can be more aware of how these things work out in the real world.