DA503 Applied Statistics

Lecture 07

Chi-square Tests

Chi square distribution

- Chi square distribution is a logical extension of basic distributions. It's a special case of the Gamma distribution.
- Denoted by χ^2, Chi square is one of the most widely used probability distributions in inferential statistics.
- Discovered twice in history, independently by K. Pearson (1900) and by F.R: Helmert (1875).
- **Chi square distribution is the distribution of a sum of the squares of k independent standard normal random variables.**
- If $Z_1, Z_2 \ldots Z_k$ are all standard normal random (iid) variables such that $Z_i \sim N(0,1)$ where $i=1,2,\ldots,k$, then (via Cochran's theorem):

$$\sum_{i=1}^{k} Z_i^{\,2} \sim \chi_k^{\,2}$$

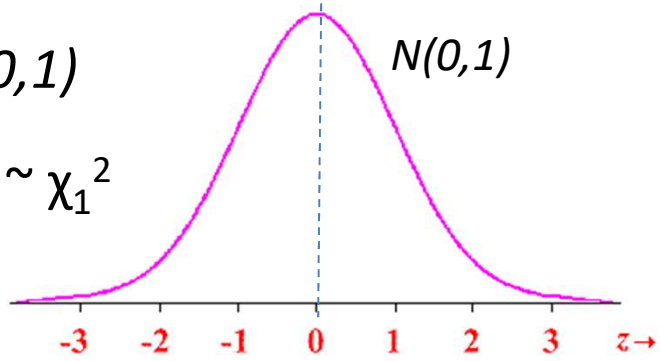
where **k** is the **degrees of freedom**

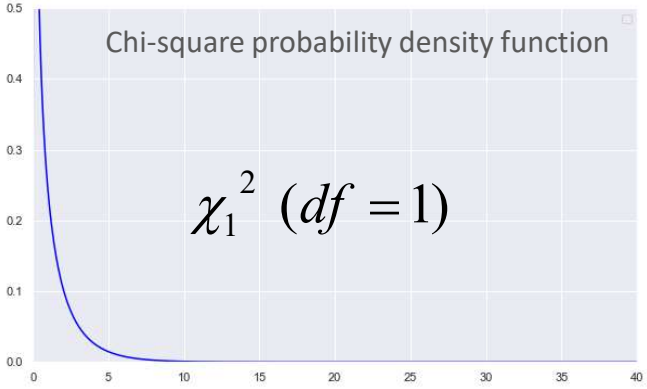
Chi square distribution

- Suppose we have a population $X \sim N(\mu, \sigma^2)$, with scores X that are normally distributed with mean μ and variance σ^2.
- If we repeatedly take samples of size n, and for each sample compute a squared standard score: $Z^2 = \left(\frac{X - \mu}{\sigma}\right)^2$
- By defining $\chi_1^2 = Z_1^2$, what would the sampling distribution of χ_1^2 (Chi square with df=1) look like?

$Z_1 \sim N(0,1)$

$Q = Z_1^2 \sim \chi_1^2$



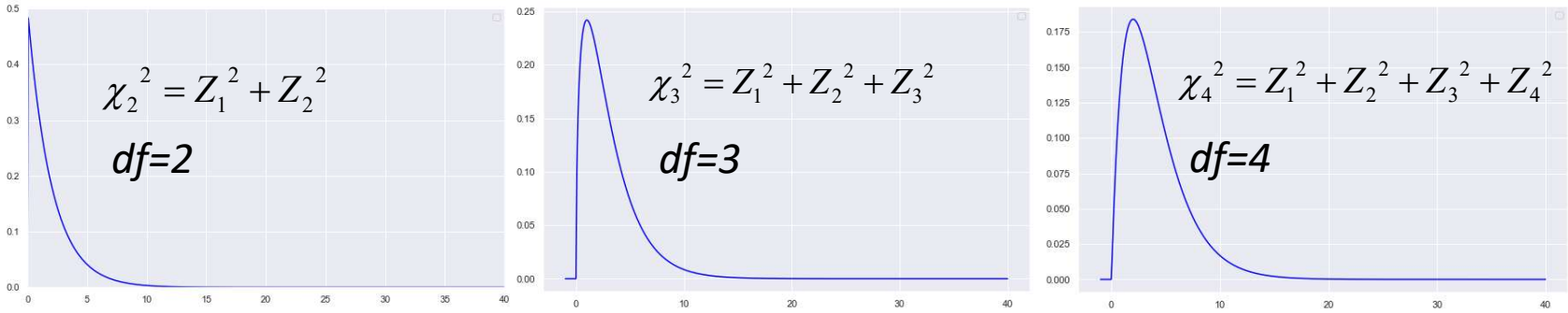


Q

As most of the values from N(0,1) are likely to come from [-1,1], χ_1^2 distribution will be even smaller with a sharp spike towards zero (χ^2 is very skewed for df=1).

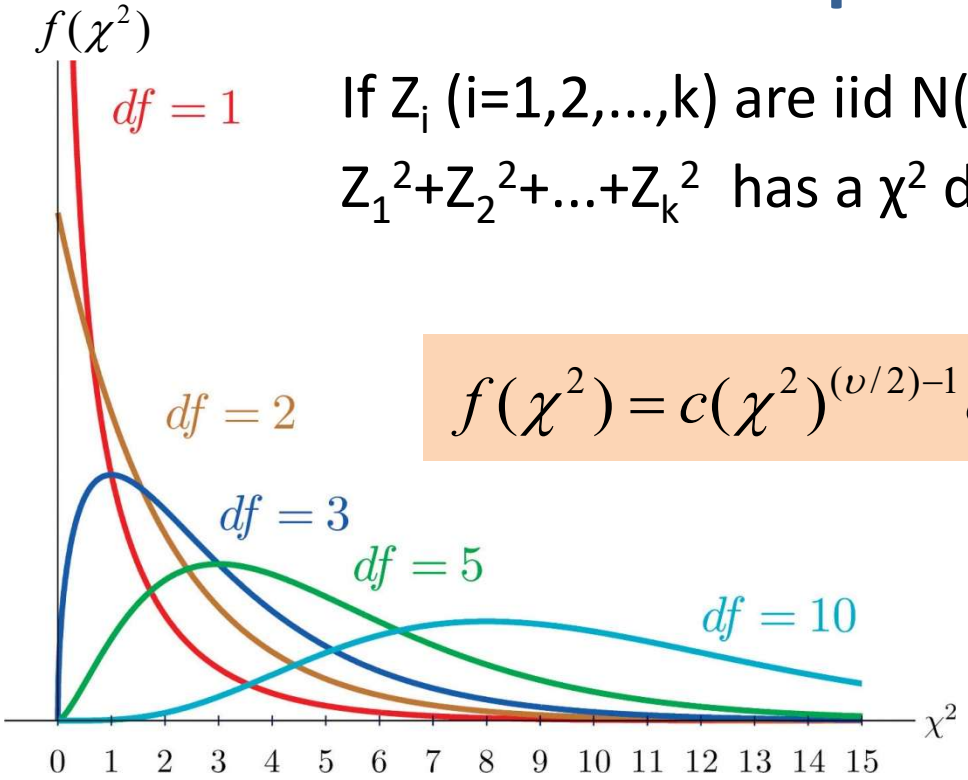
Chi square distribution – cont'd

- Repeatedly draw iid samples of size n from N(0,1) and get the sums $Z_1^2 + Z_2^2$, $Z_1^2 + Z_2^2 + Z_3^2$, and $Z_1^2 + Z_2^2 + Z_3^2 + Z_4^2$



- For df=k, the sum will be $\chi_k^2 = Z_1^2 + Z_2^2 + \ldots + Z_k^2$
- In multiple selections, less probability of selecting from the range [-1,1] moves the peak of the sum away from 0.
- A fact we'll re-visit later in this chapter:
- For samples of size n taken from a Normal distribution with variance σ^2, the sampling distribution of **$(n-1)s^2/\sigma^2$** has a Chi-square distribution with **n-1 degrees of freedom**.

Chi square distribution – cont'd



If Z_i (i=1,2,...,k) are iid N(0,1) random variables,

$Z_1^2+Z_2^2+...+Z_k^2$ has a χ^2 dist. of df=k with ($\mu=k$, $\sigma^2=2k$)

$$f(\chi^2) = c(\chi^2)^{(\upsilon/2)-1} e^{-\chi^2/2}$$

where $c = \dfrac{1}{2^{\upsilon/2}\Gamma(\upsilon/2)}$

and e=2.71828, v = df

As df increases, $f(\chi^2)$ converges to a Normal distribution

- χ^2 is a continuous distribution that depends on only one parameter called degrees of freedom (much like t-dist).
- **μ=df** , **σ^2=2df** and $f(\chi^2)$ is max when χ^2 = df-2 (for df>2)
- The random variable χ^2 varies from 0 to infinity. The coefficient c makes sure that the area underneath $f(\chi^2)$ is 1.

The Chi-square test

- Although the methodology is essentially the same, we use Chi-square test (a non-parametric test) for two tasks:
  - **Test of "Goodness of fit"**

    Goodness of fit involves a comparison of the frequency observed in the sample with the expected frequency based on some theoretical model (distribution) to determine whether the sample is taken from a population used to calculate the expected frequencies.

  - **Test of "Independence"**

    Helps us understand the relationship between 2 categorical variables. The goal of a two-variable Chi-square test is to determine whether the first variable is related (or not) to the second variable.

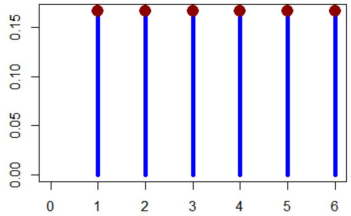
The Chi-square test – cont'd

- Example: We want to know whether a dice is fair or not. Does the collection of data from the toss of a die follow a uniform distribution with equally likely outcome for each value?(test of **goodness-of-fit**)?

  H_0: The observed pattern **fits the given distribution**

  H_A: The observed pattern does **not fit the given distribution**

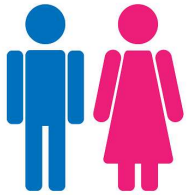
?



- Example: Is there an association (statistically significant that is) between the gender and the letter grade for a class (test of **independence**)?

  H_0: The two variables are **independent**

  H_A: The two variables are **dependent**



A, B+, C-, D, A, A+ … ?

The Chi-square test – cont'd

- Chi-square involves the frequency of events. It counts and compares observed frequencies to expected frequencies:

  χ^2 statistic used for testing: $$\chi^2 = \sum_{i=1}^{n} \frac{(o_i - e_i)^2}{e_i}$$

  o_i: observed frequency and **e_i: expected frequency**

- e_i is the frequency that would be expected in a cell, on average, if the variables are independent. χ^2 test is based on the comparison of the observed vs the expected values.
- Why square the difference and normalize by e_i?
- (o_i - e_i) is a measure of deviation from the expected frequency We're taking the square to prevent the sum from being zero.
- We want the test statistic to be independent of the spread, thus we scale it with the expected frequency. So, the scaling is there to account for the impact of large absolute values.

The Chi-square test – cont'd

- Rationale behind the χ^2 statistic?
- Suppose we want to analyze the effect of smoking on heart attack:

	Smokes	Doesn't smoke	Total
Heart attack	70	30	100
No heart attack	35	55	80
Total	105	85	

- As we work with the counts (not something continuous) and frequencies, the cell counts ($\boldsymbol{o_i}$) are independent Poisson variables when unconditioned on the total counts.

$$o_i \sim \mathbf{Poisson}(\mu_i)$$ where $\boldsymbol{\mu_i}$ is the expected value ($\boldsymbol{\lambda}$)

- Once we impose a total cell count for the table, the cell counts then become multinomial.
- In any case, for a Poisson distribution we have

$$\mathbf{E}(o_i) = \mathbf{Var}(o_i) = \mu_i$$

The Chi-square test – cont'd

- So, the standardized cell count (differenced by the expected value and normalized by the standard deviation):

$$\textbf{Standardized } o_i = \frac{o_i - \mathbf{E}(o_i)}{\sqrt{\mathbf{Var}(o_i)}} = \frac{o_i - \mu_i}{\sqrt{\mu_i}}$$

- When the expected counts are large enough, the distribution of this **standardized Poisson variable converges to a standard normal** N(0,1). Confirm this by a simulation.
- The standardized o_i is a standard normal random variable Z
- For Poisson, we know that **E**(X) = **Var**(X) = e_i (expected value)
- If we square the normalized Poisson, we get χ2:

$$\textbf{Standardized } o_i = \frac{o_i - \mu_i}{\sqrt{\mu_i}}$$

mean = e_i

std dev = $\sqrt{e_i}$

and $$Z = \frac{x - \mu}{\sigma}$$

$$\sum_{i=1}^{n} Z_i^2 \Rightarrow \sum_{i=1}^{n} \left(\frac{o_i - e_i}{\sqrt{e_i}} \right)^2 = \sum_{i=1}^{n} \frac{(o_i - e_i)^2}{e_i}$$

$$\chi^2 \sim \sum_{i=1}^{n} Z_i^2 = \sum_{i=1}^{n} \frac{(o_i - e_i)^2}{e_i}$$

The Chi-square test – cont'd

- We want to test if a die is fair (H_0: die is fair), i.e., each of the numbers {1,2,3,4,5,6} has a probability of 1/6.
- Say you toss the die 120 times and record the observed counts ($\boldsymbol{o}_i$) for i = 1,2,3,4,5,6.
- If the null is true, we would expect to see 20 (expected frequencies) for each number in 120 tosses.
- To measure how close the observed frequencies to expected frequencies, we compute the chi-square statistic:

$$\chi^2 = \frac{(o_1 - e_1)^2}{e_1} + \ldots + \frac{(o_6 - e_6)^2}{e_6}$$

a measure of deviation of the observed frequencies compared against the expected frequencies that tells us how probable these deviations are

- If $\boldsymbol{o}_i$ is close to $\boldsymbol{e}_i$, then χ^2 should be close to 0 and we would not reject H_0. Under fairly general conditions, this statistic follows what is called the chi-square distribution.

Dice experiment – Goodness of fit test

- Suppose I have two dice: one is "fair" and the other is "1-5-6 loaded" (favors 1-5-6 due to altered weight). You're given one die and asked to determine if it's the fair or the loaded one. You need to be 95% confident.
- You roll the die 600 times and record how many times each number turns up.
- If the die is fair, what would you expect to happen?
- We use chi-square distribution and critical value to accept or reject the hypothesis
- **H_0: The die is fair**

  p_i=1/6 for i=1,2,3,4,5,6

#	EXPECTED FREQUENCY*	OBSERVED FREQUENCY
1	100	**111**
2	100	**90**
3	100	**81**
4	100	**102**
5	100	**124**
6	100	**92**
TOTAL	600	600

*Expected frequency = p_i x 600

Dice experiment – Goodness of fit test

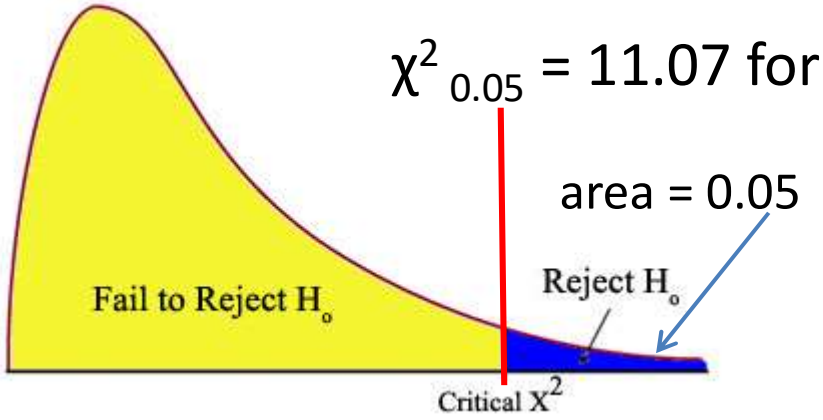
- H_0: the die is **fair** and H_A: the die is **not fair**

  Conf. level = 95% & p-value ≤ 0.05

  Degrees of freedom (df) = 6 – 1 = 5

Just 1 constraint. We already used the sum 600 to compute the expected frequencies. So only 5 of the frequencies are free to change as their sum must equal 600

$$\sum_{i=1}^{k} Z_i^2 \sim \chi_{k-1}^2$$



Chi-square Distribution Table

d.f.	.995	.99	.975	.95	.9	.1	.05
1	0.00	0.00	0.00	0.00	0.02	2.71	3.84
2	0.01	0.02	0.05	0.10	0.21	4.61	5.99
3	0.07	0.11	0.22	0.35	0.58	6.25	7.81
4	0.21	0.30	0.48	0.71	1.06	7.78	9.49
5	0.41	0.55	0.83	1.15	1.61	9.24	11.07

- If our die $\chi^2 > 11.07$, then we must reject H_0 and claim the die is not fair.

$$\chi^2 = \sum_{i=1}^{6} \frac{(o_i - e_i)^2}{e_i} = 1.21 + 1 + 3.61 + 0.04 + 5.76 + 0.64 = 12.26$$

$\chi^2 = 12.26 > 11.07$, so the die is **NO**T fair (H_0 rejected)

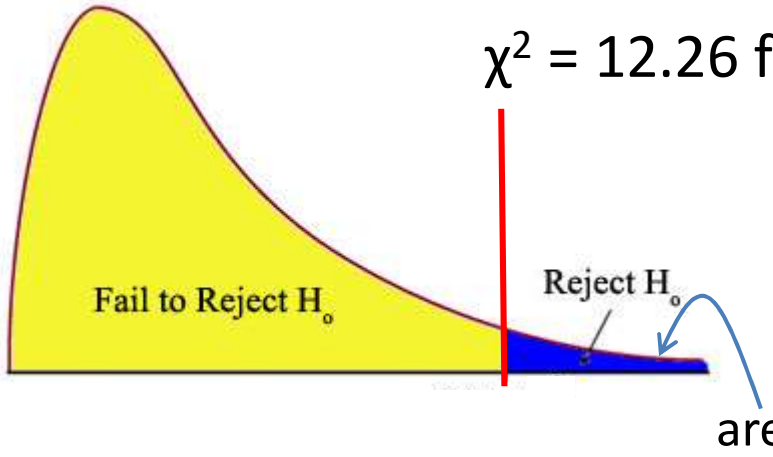
Dice experiment – Goodness of fit test

- Solution via Python:

Python code

```
import scipy.stats as stats
observed = [111,90,81,102,124,92]
chi2, p = stats.chisquare( observed )
msg = "Test Statistic: {}\np-value: {}"
print( msg.format( chi2, p ) )
Test Statistic: 12.260000000000002
p-value: 0.031393731655486354
```

- χ^2 is 12.26 (as computed previously) and the probability $p(\chi^2 > 12.26) = 0.031$ is less than 0.05. So, we reject H_0.



$$\Pr(\chi^2 > 12.26) = 0.031$$

```
1-stats.chi2.cdf(12.26,df=5)
>>> 0.03139373
```

Dice experiment – Goodness of fit test

- **A note on the use of one-sided vs two-sided χ^2 tests**
- If you're testing variance of normal data against a specified value, you might be dealing with the upper or lower tails of the chi-square.
- A goodness-of-fit test, however, is essentially *always a one-sided test*. When the computed χ^2 value is way out on the right tail of its distribution, it indicates a poor fit, and if it is far enough, relative to some pre-specified threshold, we might conclude that it is so poor that we don't believe the data are from that reference distribution.
- If we were to use the χ^2 test as a two-sided test, that means we also care about if the statistic were too far into the left side of the distribution which could mean that the fit might be *too good*. This is something we typically don't care about.

Source: stats.stackexchange.com/questions/22347/is-chi-squared-always-a-one-sided-test

Goodness of fit test

- **Example 1**: You went on camping with your best friends and ended up with the following table about who did the dishes?

You	Alice	Bob	John	Mary	David
10	6	5	4	5	3

- You kind of suspect that this wasn't exactly fair. How could you put this to a test?
- Expected frequency = N_{total} / N_{people} = 33/6
- Is there a significant difference between the observed and expected frequencies?

```
data = [10, 6, 5, 4, 5, 3]
chi2, p = stats.chisquare(data)
print(p)
>>> 0.373130385949
```

- Not enough evidence to claim that you've done most of the dishes.

Goodness of fit test

- **Example 2**: Chi square goodness-of-fit test with unequal expected frequencies:

Reports show that 40% of the students never visit infirmaries, 30% once a year and so on. The local school wants to know how its own records compare with the national pattern with 95% confidence.

Number of times admitted	Observed number of admissions	Expected percent - (national)
0	**55**	40%
1	**50**	30%
2	**32**	20%
3 or more	**13**	10%
Total	150	100%

Observed freq (o_i)	Expected freq (e_i)
55	150*0.40 = **60**
50	150*0.30 = **45**
32	150*0.20 = **30**
13	150*0.10 = **15**

H_0: p_1, p_2, p_3, p_4 are equal to hypothesized proportions (no difference between local and national patterns.

Test statistic : $\chi^2 = \sum_{i=1}^{4} \frac{(o_i - e_i)^2}{e_i} = 1.372$

Critical value $\chi^2_{0.05, df=4-1=3} = 7.815$

Failed to reject H_0 (no difference between local & national records)

Chi-square test for normality

- Chi-square goodness-of-fit test can also be used to test whether or not a sample comes from a normal distribution
- Example: Given the following sample (**n = 40**), find out if it comes from a normal distributed population.

14.92	16.76	12.61	11.95	12.74	10.37	11.18	14.06
11.95	14.12	14.10	15.78	16.72	13.95	11.30	11.62
14.37	11.74	14.54	13.21	16.44	16.35	13.00	16.68
16.05	15.29	15.30	15.51	16.04	12.22	16.03	11.03
11.22	10.42	14.54	12.11	18.53	14.74	11.27	16.21

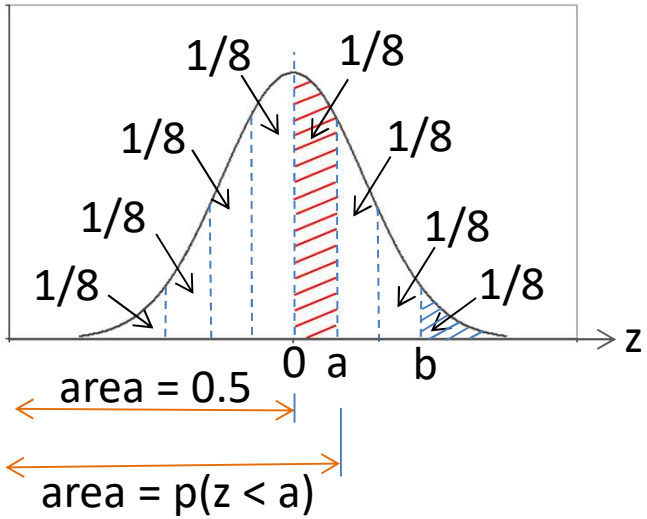
H_0: The population probability distribution is normal (all proportions are equal to a hypothesized value)

H_A: Non-normal (at least one proportion differs)

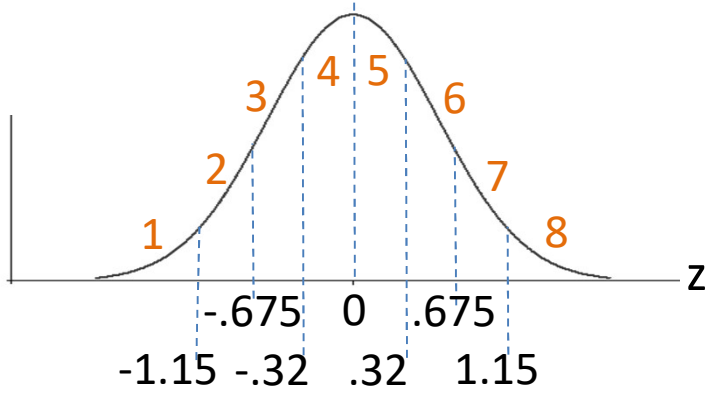
- Let's divide the normal distribution into **8 bins** each of which **with a probability of 1/8**.

From: www.youtube.com/watch?v=4SAqlgfWmjA

Chi-square test for normality



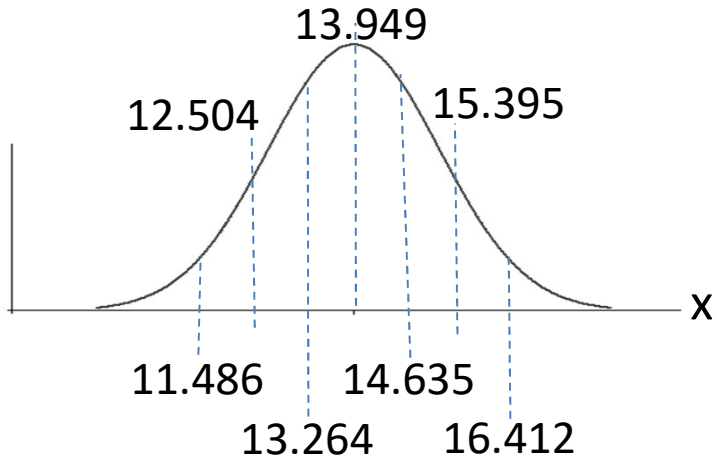
shaded area (blue) = 1 – p(z < b) = 1/8
shaded area (red) = p(z < a) – 0.5 = 1/8
From the look-up table (or via Python), we find b = 1.15 and a = 0.32. You get all z values for each bin in a similar fashion



- Estimated values of the sample mean and variance are:

$$\bar{x} = 13.949 \qquad S = 2.1416$$

- Compute the x values from the z scores above: $z = \dfrac{x - \bar{x}}{S}$



Chi-square test for normality

- We sort the data in the ascending order:

10.37 10.42 11.03 11.18 11.27 11.30 11.62 11.72 11.74 11.95
11.95 12.11 12.22 12.61 12.74 13.00 13.21 13.95 14.06 14.10
14.12 14.37 14.54 14.54 14.74 14.92 15.29 15.30 15.51 15.78
16.03 16.04 16.05 16.21 16.35 16.68 16.72 16.76 16.94 18.53

interval	observed	expected
≤ 11.486	6	5
11.486 to 12.504	7	5
12.504 to 13.264	4	5
13.264 to 13.949	0	5
13.949 to 14.635	7	5
14.635 to 15.395	4	5
15.395 to 16.412	7	5
≥ 16.412	5	5

Expected value:
$e_i = np = 40 * 1/8 = 5$

Chi-square test for normality

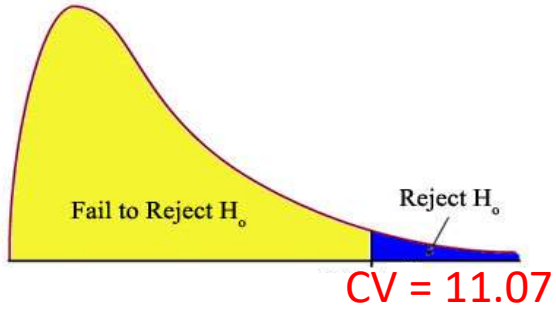
- **H_0: Normal distribution** and **H_A: Non-normal**
- Degrees of freedom: $df = k - 1 - m$

number of constraints imposed on the expected frequencies (total = 40 is known)

m : number of population parameters estimated from the data (sample mean + variance = 2) from the data

df = 8 – 1 – 2 = 5 (if the population mean & variance were known, df would be just 7)

$$\chi^2 = \sum_{i=1}^{8} \frac{(o_i - e_i)^2}{e_i} = 8 \quad \textbf{and} \quad \textbf{df} = \textbf{5}$$



- For α=0.05 and df = 5, we find the critical value as 11.07 and since 8 < CV, we have insufficient evidence to reject the Null hypothesis which claimed normality. So, the **sample comes from a normally distributed population**.

Chi-square test of independence

- One of the most basic and common hypothesis tests in the statistical analysis of categorical variables. This test determines whether there exists a statistically significant dependency (association) between them.
  - H_0: Variables are independent (no relationship)
  - H_A: There is a significant relationship between variables
- Example: We try to determine whether there is a dependency between gender and the preference for coffee type? Is there a significant difference in preference for coffee type between men and women?

Gender has 2 categories	**Coffee** has 3 categories
• **Men** • **Women**	• **Espresso** • **Cafe Latte** • **Cappuccino**

- In a survey, 880 people are asked for their preferences:

Example

2-dimensional ***contingency table***		Type of coffee drink			
		Latte	**Espresso**	**Cappuccino**	Total
Gender	**Men**	80 **(105)**	80 **(75)**	60 **(40)**	220 **(25%)**
	Women	340 **(315)**	220 **(225)**	100 **(120)**	660 **(75%)**
	Total	420	300	160	880

- Men form 25% (220/880) of all human subjects. So, the expected value for men who are Latte drinkers would be: 0.25 x 420 = 105. Similarly, the expected value for women who are Latte drinkers is: 0.75 x 420 = 315. Upon carrying out the same calculation and filling out the expected frequencies for all the other cells (numbers in red), we get the χ^2 value:

$$\chi^2 = \sum_{i=1}^{6} \frac{(o_i - e_i)^2}{e_i} = (80-105)^2/105 + (80-75)^2/75 + (60-40)^2/40 + (340-315)^2/315 + (220-225)^2/225 + (100-120)^2/120 = 21.71$$

Determining the Degrees of Freedom

- Degrees of freedom in a goodness of fit test:
  - We always subtract at least 1 one from K (the number of categories) because of the constraint that the expected frequencies must sum to n (total number of frequencies in all categories). Thus, if K-1 expected frequencies are known, the remaining frequency is determined.
  - df = K -1 if the expected frequencies are computed without estimating any population parameters from the sample data
  - df = K – 1 – r if the expected frequencies are computed only after estimating r population parameters from the sample data
- Degrees of freedom in a test of independence:

For R rows and C columns in a contingency table,
the number of degrees of freedom is: df = (**R**-1) (**C**-1)

Determining the Degrees of Freedom

- Degrees of freedom for contingency tables
  - For a 2 row - 3 column contingency table:, df = (**R**-1) (**C**-1) = 2

	C1	C2	C3	
R1	empty	empty	empty	**Total**
R2	empty	empty	empty	**Total**
	Total	**Total**	**Total**	

Suppose we have a two-way table with two categorical variables. One has three levels and the other has two.

Constraints imposed by the number of observations

	C1	C2	C3	
R1	80	empty	empty	200
R2	20	empty	empty	400
	100	200	300	

Suppose that we fill in the upper left cell with the number 80. This will automatically determine the other entry in the first column.

	C1	C2	C3	
R1	80	50	empty	200
R2	20	empty	empty	400
	100	200	300	

Further suppose that we fill the entry in the 2nd column from the same row with 50, then the rest of the table can be filled in as we know the total in each row and column. So, **we only have 2 free choices**.

Example – Python code and the output

```
table_ = [ [ 80,80,60 ], [ 340,220,100 ] ]
chi2s, p, ddof, expected = scipy.stats.chi2_contingency(table_)
msg = "Test Statistic: {}\np-value: {}\nDegrees of Freedom: {}\n"
print( msg.format( chi2s, p, ddof ) )
```

```
Test Statistic: 21.7142857
p-value: 1.92665e-05
Degrees of Freedom: 2
```

```
1-stats.chi2.cdf(21.714, df=2)
>>> 1.926924994022361e-05
```

- Degrees of freedom: df = (**C** - 1)x(**R** – 1) = (**3** - 1)x(**2** - 1) = 2
- $p < 0.05$ suggests that gender and coffee type preferences are dependent or have some association (reject the Null).
- Practical implication: We have strong evidence to believe that men and women tend to have different preferences for the type of coffee drink. The alternative hyp doesn't specify the type of association, so close attention to data is required to interpret the information provided by the test.

Post-Hoc testing

- Chi-square is an Omnibus test which tests for an overall dependency between nominal variables: Is there a significant association between variables? If so, it doesn't tell us which components contribute to significance.
- When the chi-square test produces a significant result (it's likely that the two variables are dependent), we may want to know which components of the two variables are responsible for the implied association.
- We can conduct additional chi-square tests based on a subset (pairwise) of the original contingency table. This is called a **post-hoc test**.
- This technique, however, creates an (familywise) error resulting from the multiple chi-square tests conducted on several pairs.

Post Hoc testing

- So, we follow a significant Chi-square test with a post-hoc (meaning after) test. There are multiple ways we can conduct a **post-hoc test**.
- Simplest is the **Bonferroni** (correction) **test** which is a series of χ^2 square tests performed on each pair of components.
- Consider a case where you have 10 hypotheses to test, and a significance level of 0.05. What's the probability of observing at least one significant result just due to random sampling variability?

P^* = P(at least one significant result due to sampling variability)

P^* = 1 – P(no significant results) = $1 - (1-0.05)^{10} \approx 0.40$

- So, with 10 tests being considered, we have a 40% chance of observing at least one significant result, even if all the tests are indeed not significant.

Post Hoc testing

- So, we need to adjust **α** in some way, so that the probability of observing at least one significant result due to chance remains below our desired significance level (**α**).
  - For **k** groups, there are C(k,2) combinations of different pairs. We want to test them all with α_c where

$$\alpha_c = \alpha/C(k,2)$$

  - If we have 5 groups, for example, we end up testing C(5,2)=10 pairs with a corrected level of significance α_c:

For α=0.05 and 10 hypotheses, $\alpha_c = 0.05/10 = 0.005$

P^* = P(at least one significant result due to sampling variability)

P^* = 1 – P(no significant results) = $1 - (1-0.005)^{10} \approx 0.049$

Slightly less than 0.05, so Bonferroni correction is a little conservative.

Post Hoc testing

- For the Gender-Coffee type problem we analyzed earlier, we found a significant dependency between the two.
- Which categories are contributing to this significance?
- **Solution:**
- Step 1: Construct pairwise combinations of Coffee type
- Step 2: Conduct a Chi-square test on pairwise combinations of Coffee types vs Gender
- Step 3: Find the p-values for each test
- Step 4: Compare these values with the Bonferroni corrected alpha
- Step 5: Decide which categories of Coffee contribute to significant relationship between Gender and Coffee type.
- Next slide for results (see the notebook for the code).

Chi square residuals

- A (Pearson) residual is the **difference between the observed and expected values for** a cell normalized by the square root of the expected value. The larger the residual, the greater the contribution of the cell to the magnitude of the computed χ^2 statistic:

$$r_{ij} = (o_{ij} - e_{ij}) / \sqrt{e_{ij}} \qquad \text{Note that} \quad \sum_{i,j} r_{ij}^{\,2} = \chi^2$$

- Under H_0, $\boldsymbol{r_{ij}}$ are asymptotically normal with 0 mean. The variance for $\boldsymbol{r_{ij}}$, however, is less than 1. To compensate for this, we use the standardized (adjusted) Pearson residuals:

$$(r_{ij})_s = \frac{o_{ij} - e_{ij}}{\sqrt{e_{ij}(1 - p_{i.})(1 - p_{.j})}}$$

Asymptotically distributed as a standard normal

where $p_{i.} = n_i / N$ **and** $p_{.j} = n_j / N$ are the estimated row i and column j marginal probabilities, and N is the sample size.

Post-hoc test and the residuals

- We can use the post-hoc tests and the residuals to identify the components that contribute most to the overall significance in a contingency table.
- **Post-hoc test** for the Gender-Coffee problem

pairs	uncorrected p	corrected p	reject?
(cappuccino, espresso)	0.021524	0.064572	False
(cappuccino, latte)	0.000006	0.000017	True
(espresso, latte)	0.019624	0.058873	False

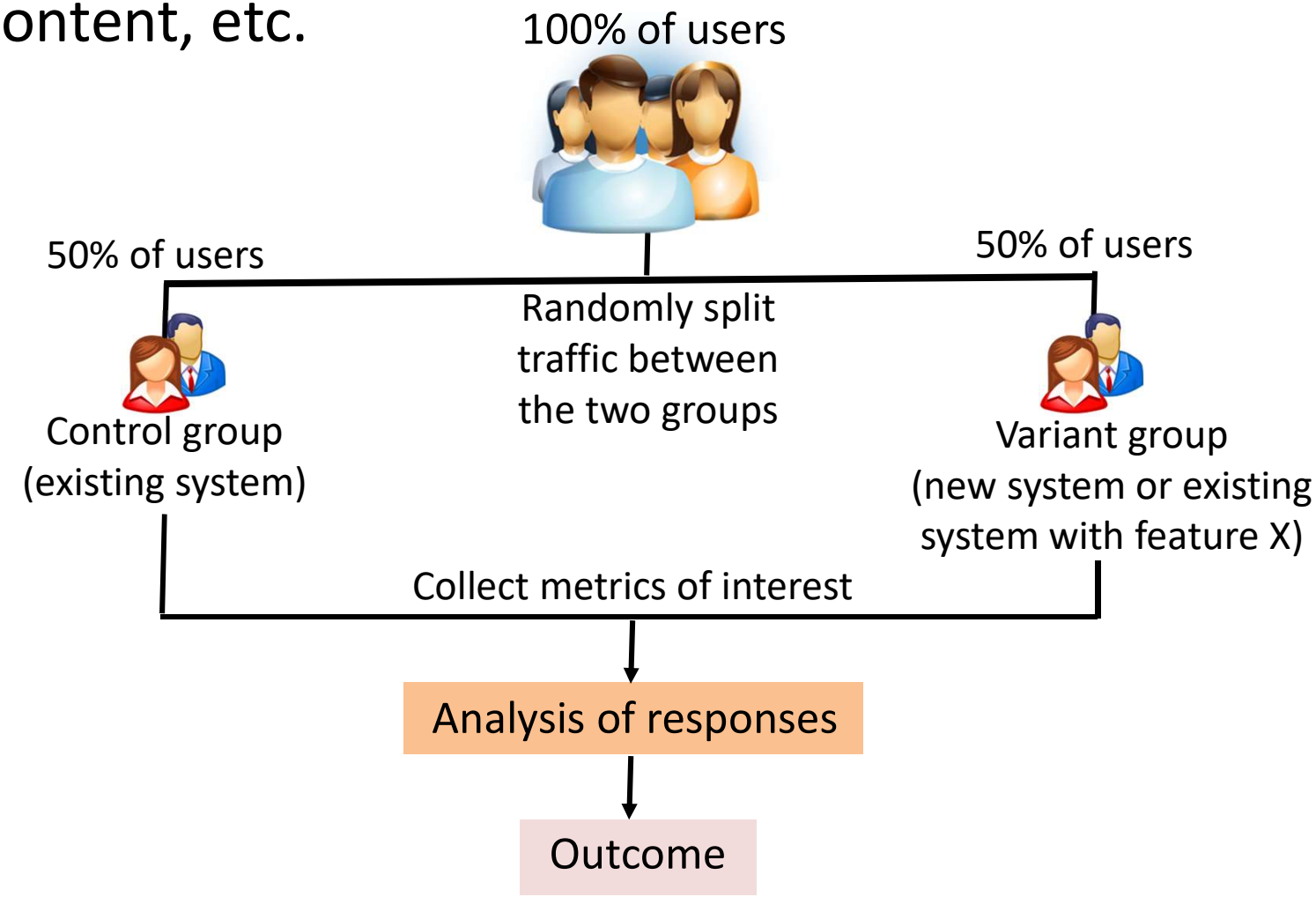
- **Standardized residuals:**

p values computed from z scores:

drink	gender	pval
male	latte	0.000293
male	espresso	1.000000
male	cappuccino	0.000163
female	latte	0.000293
female	espresso	1.000000
female	cappuccino	0.000163

Bonus example: A/B Testing

- Used for testing landing page design (web), conversion rates on different ads, e-store checkout page, email content, etc.



A/B Testing – cont'd

- Using A/B split testing to reduce bounce rate for an e-commerce store

In-your-face 'Holiday Sale' message displayed in big, red font prominently on the homepage

Sidebar 'Holiday Sale' message in small font

A/B Testing – cont'd

Message location	Impressions	Clicks (conversions)	Non clicks	CTR	Bounce rate
Sidebar	14,000	105	13,895	0.75%	99.25%
In-your-face	19,500	110	19,390	0.56%	99.44%

Reduction in bounce rate

H_0: There is no difference in conversion rates

H_A: There is evidence of a significant difference

```
abTest = [ [13895,105], [19390,110] ]
chi2s,p,ddof,expected = scipy.stats.chi2_contingency(abTest)
msg = "Test Statistic: {}\np-value: {} \nExpected \
frequencies: \n {}"
print( msg.format(chi2s, p, expected))
```

```
Test Statistic: 4.12967386
p-value: 0.04213747
Expected frequencies: [[13910.14925373    89.85074627]
                       [19374.85074627   125.14925373]]
```

Reject H_0: **statistically significant** increase in the conversion rate

- How about the effect size? See Lecture-08 on correlations.

A/B Testing – cont'd

- This problem could've been solved using a z-test for the difference of proportions where H_0: p_1-p_2=0

$$Z = \frac{(\hat{p}_1 - \hat{p}_2)}{\sqrt{(\hat{p}\hat{q}/n_1)+(\hat{p}\hat{q}/n_2)}} \quad \text{where} \quad \hat{p} = \frac{X_1 + X_2}{n_1 + n_2} = \frac{105+110}{14000+19500} \quad , \quad \hat{q} = 1 - \hat{p}$$

$$\hat{p}_1 = 105/14000 \quad , \quad \hat{p}_2 = 110/19500$$

- This gives a z-statistic: `z = 2.10152042`

```
from statsmodels.stats.proportion import proportions_ztest
total = np.array([105, 110])
clicks = np.array([14000, 19500])
z,p = proportions_ztest(total,clicks,alternative='two-sided')
print('zscore = {:.6f}, pvalue = {:.6f}'.format(z, p)
 zscore = 2.101520, pvalue = 0.035595
```

So, the increase in conversion rate is significant (Reject H_0)

How does this compare with the χ^2 (previous slide)?

```
Chi2, p, _, _ = chi2_contingency(abTest, correction=False)
print(p)
0.035595306679237 3
```

set to **False** to avoid Yates' correction (later)

A/B Testing – cont'd

- Note that the chi-square statistic found earlier is the square of the 2-tailed 1-sample proportion Z statistic.
- Z statistic: z = 2.101 and χ^2 statistic = 4.414 = z^2

	Gr1	Gr2	Total
Yes	p1	p2	p
No	q1	q2	q
	100%	100%	100%
	n1	n2	N

The usual (not Yates corrected) χ^2 of this table, after you substitute proportions instead of frequencies in its formula, looks like this:

$$n_1[\frac{(p_1-p)^2}{p}+\frac{(q_1-q)^2}{q}]+n_2[\frac{(p_2-p)^2}{p}+\frac{(q_2-q)^2}{q}]=\frac{n_1(p_1-p)^2+n_2(p_2-p)^2}{pq}$$

Remember that $p=\frac{n_1p_1+n_2p_2}{n_1+n_2}$, the element of the weighted average profile of the two profiles `(p1,q1)` and `(p2,q2)`, and plug it in the formula, to obtain

$$\ldots=\frac{(p_1-p_2)^2(n_1^2n_2+n_1n_2^2)}{pqN^2}$$

Divide both numerator and denominator by the $(n_1^2n_2+n_1n_2^2)$ and get

$$\frac{(p_1-p_2)^2}{pq(1/n_1+1/n_2)}=Z^2,$$

Ref: stats.stackexchange.com/questions/173415/at-what-level-is-a-chi2-test-mathematically-identical-to-a-z-test-of-propo

A/B Testing – cont'd

- Can we address this A/B test by simulation?

```
import random
cA = 105 ; ncA = 13895
cB = 110 ; ncB =  19390
import random
total = cA + ncA + cB + ncB
totA = cA + ncA
totB = cB + ncB
ratio_diff = (cA/(cA+ncA)) - (cB/(cB+ncB))
```

```
# cliks are represented by 1's -- we have a total of cA + cB 1's
ones = np.ones(cA + cB)
zeros = np.zeros(ncA + ncB)
pool = ones.tolist() + zeros.tolist()
```

```
iters = 50000
counter = 0
for i in range(iters):
    random.shuffle(pool)
    listA = pool[:totA]
    listB = pool[totA:]
    clickA = sum(listA)
    clickB = sum(listB)
    delta = (clickA/(clickA+(len(listA)-clickA))) -
            (clickB/(clickB+(len(listB)-clickB)))
    if abs(delta) > abs(ratio_diff): counter += 1
print(counter/iters)
```

```
0.03266
```

<= compares well with our previous z-test and chisquare-test

Tips on A/B tests

- As a rule of thumb:
  - Confidence intervals in A/B testing: Get as close to 99% confidence level as possible
  - Sample size in A/B testing: At least 100 conversions
  - Conversion range less than +/- 1% of standard error
- Things to watch out in A/B testings:
  - Avoid the traps of misinterpretations (like p-values)
  - Avoid significance peeking (watching the p-values as the test is on and stopping early thinking that we reached a statistically significant result). There may be fluctuations early on in the tests leading you to believe that you reached significance.
  - Avoid non-simultaneous testings (beware of day of week effects). Both the control and variant groups should be tested at the same time to prevent time-induced effects.

Limitations on Chi-square test

- Check these 2 conditions before performing a χ^2 test:
- **Independence**: Each case that contributes a count to the table must be independent of all the other cases.
- **Sample size / distribution**:
  - For a 2x2 table, all expected frequencies > 5
  - For larger tables, all expected frequencies > 1, and no more than 20% of all expected frequencies < 5
- Generally speaking, when these conditions on the counts of the expected freq's are not met, the asymptotic justification of the χ^2 test may not be appropriate (risk for Type I error). In such cases:
  - Collect more data (increase sample size if possible)
  - Merge categories with expected frequency counts less than 5 (re-binning the cells that are of similar nature
  - Remove categories with expected values less than 5
  - Use **Yates' correction** if a 2x2 table (see next page)
  - Use **Fisher's Exact Test** (later)

Yates' correction

- Yates' correction for 2x2 contingency tables
  - This correction modifies the χ^2 statistic for 2x2 contingency table to correct the error made by using a continuous χ^2 distribution to approximate the observed discrete (dichotomous) sampling distribution of the statistic. It's been argued that the uncorrected results are biased upwards for 2x2 contingency tables.
  - Yates' correction subtracts ½ from the size of each residual and the formula becomes:

$$\chi^2_{corrected} = \sum_{i=1}^{6} \frac{(|o_i - e_i| - 0.5)^2}{e_i}$$

  - As a side note, Monte Carlo simulation research published so far, however, recommends not using Yates' correction due to its over-corrective behavior.
  - Yates correction is applied by default for 2x2 contingency tables. The correction option must be used to turn it off:

```
chi2_contingency(table, correction=False)
```

Fisher's exact test

- The χ^2 test can be used (a standard rule of thumb) for tables larger than 2x2 and it's quite valid if:
  - Each observation is independent of all others (one observation per subject)
  - Sample size is large (no more than 20% of the expected counts are ≤ 5 and all individual expected counts are 1 or greater)
- Sometimes it's appropriate to group certain categories to avoid the problem, but this is clearly not possible when there are only two categories.
- Use **Fisher's exact test** if expected count is ≤ 5 in one or more of the cells in a 2x2 contingency table.
- Fisher's test is exact because it uses the exact hypergeometric distribution rather than the approximate χ^2 distribution to compute the p-value.

Fisher's exact test – cont'd

- It’s appropriate to use Fisher’s exact test when dealing with small counts. χ^2 test is basically an approximation of the exact test, so erroneous results could potentially be obtained from the few observations.
- **Example**: Paul and David try to guess which one is heavier when they're shown 2 items. Assuming each item has the same probability of being guessed, is there a difference between Paul's and David's guesses?

	Correct guess	**Incorrect guess**	TOTAL
David	**8** (6)	**1** (3) ?	9 (50%)
Paul	**4** (6)	**5** (3) ?	9 (50%)
TOTAL	12	6	18

Fisher's exact test – cont'd

- Here are the results from the χ^2 test:

```
>>> table_ = [ [ 8,1 ], [ 4,5 ] ]
>>> res = chi2_contingency(table_, correction = False)
>>> print(res[1])
    0.0455
```

`0.133` with correction =`True`

- p-value, 0.046, from the χ^2 test indicates there is a statistically significant difference (at the $\alpha = 0.05$ level) in the success rates between David and Paul.
- Output of the Fisher's exact test (for 2x2):

```
>>> oddsratio , pvalue = fisher_exact([[ 8,1 ], [ 4,5 ]])
>>> print(pvalue)
    0.131
```

- With p=0.13, we fail to reject the Null, which indicates that the χ^2 test above without Yates' correction provided a poor approximation to the exact results.

Fisher's exact test – cont'd

- Hand calculation of Fisher's exact test is extremely tedious and time consuming. It's a lot faster and easier by use of statistical packages now and Fisher's exact test could be used in place of the χ^2 test.
- Fisher's exact test could also be extended beyond 2x2 contingency tests. Python doesn't have this functionality, but the R function can be called from within Python:

```
>>> import rpy2.robjects.numpy2ri
>>> from rpy2.robjects.packages import importr
>>> rpy2.robjects.numpy2ri.activate()
>>> stats = importr('stats')
>>> m = np.array([[4,4],[4,5],[10,6]]) # 2x3 table
>>> res = stats.fisher_test(m)
>>> print('p-value: {}'.format(res[0][0]))
```

McNemar's test

- This is not testing for independence, but consistency in responses across two variables. It's generally used for **repeated measures** or **paired nominal data situations**.

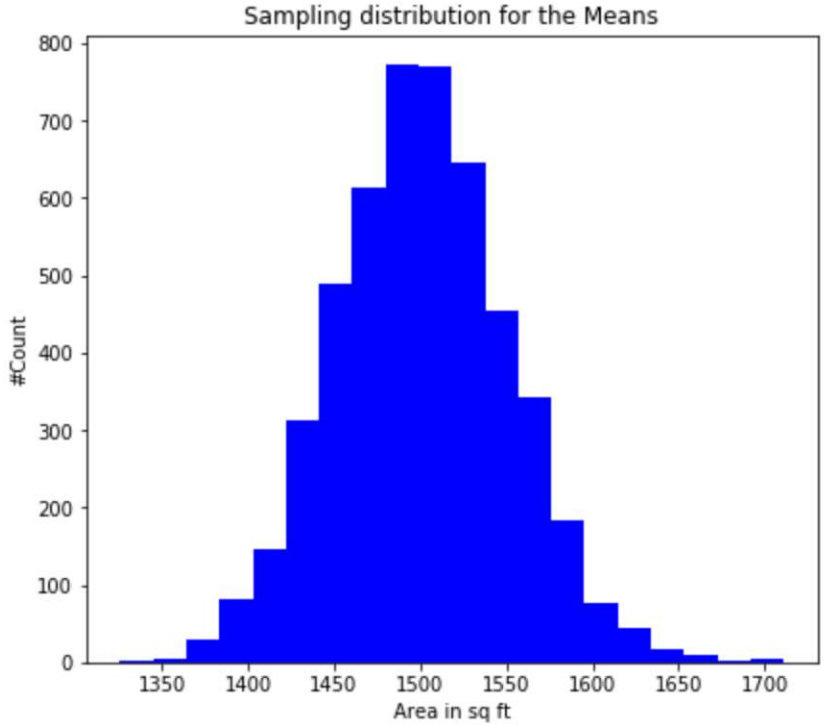
		Experience pain after treatment?	
		Yes	**No**
Experience pain before treatment?	**Yes**	a	b
	No	c	d

- If the treatment is having no effect, the number of people who move from No to Yes should be about equal to those who move in the other direction.
- In the McNemar test, we can compare counts directly, because the comparison is not based on row totals.
- McNemar test is used for 2x2 contingency tables only.

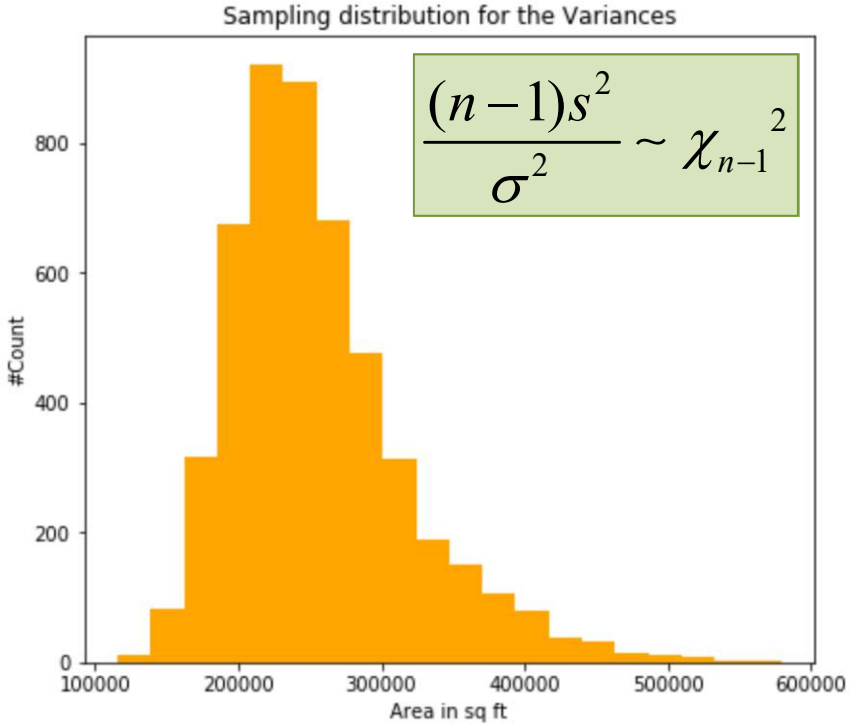
```
from statsmodels.stats.contingency_tables import mcnemar
print(mcnemar(data, exact=False)) # exact=Yates' correction
```

Confidence interval for the variance

- CLT: Sampling distribution for the means vs variances
- What happens when you take many samples from a normally distributed population and plot the sampling distribution for the sample variance?



Follows a normal distribution



Follows a χ^2_{n-1} distribution

Confidence interval for the variance

- χ^2 distribution: sum of the squares of k independent , identically distributed std normal random variables Z

$$\sum_{i=1}^{k} Z_i^{\,2} \sim \chi_k^{\,2}$$

$$\text{Given} \quad Z = \frac{x-\mu}{\sigma} \quad \Rightarrow \sum_{i=1}^{k} Z_i^{\,2} = \sum_{i=1}^{k} \left(\frac{x_i - \mu}{\sigma} \right)^2 = \frac{\sum_{i=1}^{k} (x_i - \mu)^2}{\sigma^2} \sim \chi_k^{\,2}$$

$$\sum_{i=1}^{k} (x_i - \bar{x})^2 \sim \sigma^2 \chi^2_{k-1}$$

minus 1 df

$$s^2 = \frac{\sum_{i=1}^{n} (x - \bar{x})^2}{n-1} \quad \Rightarrow \quad (n-1)s^2 = \sum_{i=1}^{n} (x - \bar{x})^2 \sim \sigma^2 \chi^2_{n-1}$$

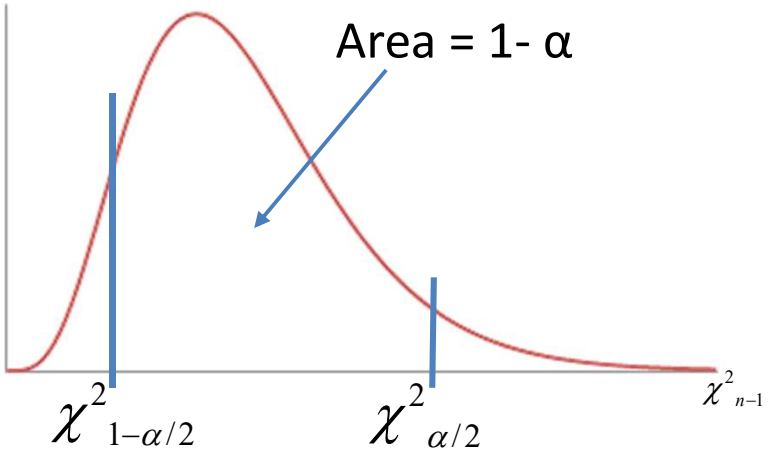
- **$(n-1)s^2/\sigma^2$** is distributed according to a chi squared distribution with **n-1** degrees of freedom:

$$\frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{n-1}$$

Confidence interval for the variance

- To construct a confidence interval for the population variance for a significance level of α

$$p\left(\chi^2_{1-\alpha/2} \le \frac{(n-1)s^2}{\sigma^2} \le \chi^2_{\alpha/2} \right) = 1-\alpha$$



We use s^2 to estimate σ^2:

$$\left(\frac{(n-1)s^2}{\chi^2_{\alpha/2}} \le \sigma^2 \le \frac{(n-1)s^2}{\chi^2_{1-\alpha/2}} \right)$$

- **Example**: A sample of 10 observations from a normal dist. has a sample variance 50. 90% confidence interval for the true population variance?

$$\left(\frac{9*50}{16.919} \le \sigma^2 \le \frac{9*50}{3.325} \right)$$

$$= (26.6 \le \sigma^2 \le 135.33)$$

$$\chi^2_{\alpha/2,df=9} = 16.919$$

```
stats.chi2.ppf(0.95, df=9)
```

$$\chi^2_{1-\alpha/2,df=9} = 3.325$$

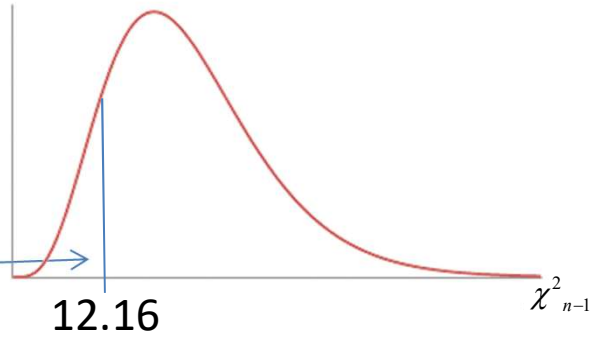
```
stats.chi2.ppf(0.55, df=9)
```

Example

- My commute time to work is normally distributed with a mean of 40 minutes and variance of 100 minutes^2. A new route I found recently seems to yield less variability.
- So with a random sample of 20 commutes to work using this new route, sample variance is down to 64 minutes^2. Is there evidence that this is really less variable with α=0.1?
- Commute time: C ~ N(40,10)
- H_0: σ^2 =100 min^2 and H_A: σ^2 < 100 min^2

$$\chi^2_{n-1} = \frac{(n-1)s^2}{\sigma^2} = \frac{19*64}{100} = 12.16$$

```
stats.chi2.cdf(12.16,19)
0.121327224246
```



- As p=0.12 > 0.10, we fail to reject H_0. New route is not less variable.

Caveats

- Be careful when constructing your categories for the tables. A chi-square test will give you information based on how you divide up your data. It cannot tell you whether your data construction makes sense or not.
- A chi-square test is meant to test the probability of independence given the distribution of data. It won't give you any details about the relationship between them.
- The variables used in the test must be mutually exclusive (participation in one category should not allow participation in another – no double-counting an item)
- Never exclude some portion of your data set as it might contain a critical part of the process.