The χ^2 test

Arnaud Legrand and Jean-Marc Vincent

Scientific Methodology and Performance Evaluation

Definition

Origin:

- Measurement errors are typically distributed with a normal distribution.
- The larger the error, the higher the risk. Let's consider the square of errors and sum them over k measurements.

Let's consider k independant variables $X_1, \ldots, X_k \sim \mathcal{N}(0, 1)$. Then:

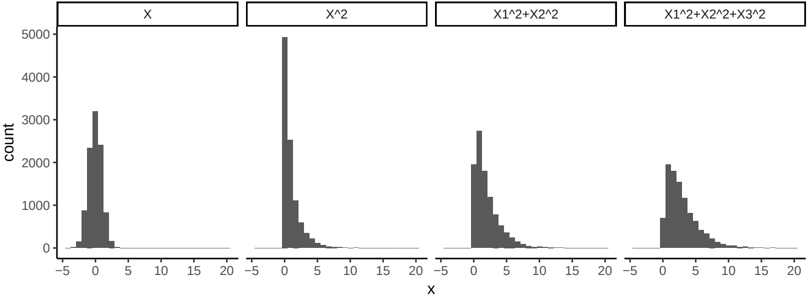
$$Q = \sum_{i=1}^{k} {X_i}^2$$

is distributed according to the χ^2 distribution with k degrees of freedom ($Q \sim \chi^2_k$).

The χ^2 distribution has one parameter: $k \in \mathbb{N}^*$ that specifies the number of degrees of freedom.

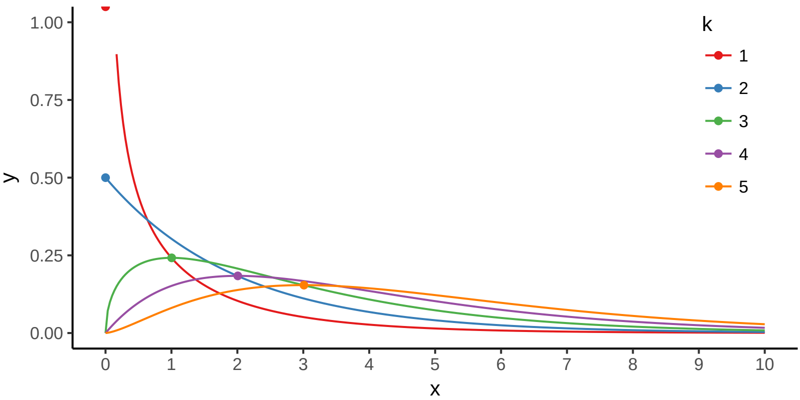
Sample Histograms

```
N=10000;
X0 = rnorm(N);                    X1s = rnorm(N)**2;
X2s = X1s + rnorm(N)**2;          X3s = X2s + rnorm(N)**2;
df=rbind(data.frame(x=X0,lab="X"),data.frame(x=X1s,lab="X^2"),
         data.frame(x=X2s,lab="X1^2+X2^2"),
         data.frame(x=X3s,lab="X1^2+X2^2+X3^2"))
ggplot(data=df, aes(x=x)) + geom_histogram() +
    facet_wrap(~lab, nrow=1) + theme_classic()
```



Probability distribution

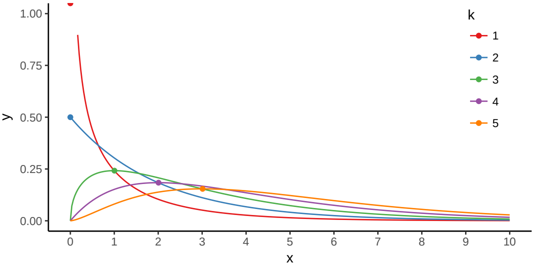
- Density function: $\dfrac{1}{2^{k/2}\Gamma(k/2)}\, x^{\frac{k}{2}-1}e^{-\frac{x}{2}}$



- As k increases, the distribution gets more and more flat and moves to the right.

Main Characteristics

- Asymetrical
- Mode at $k-2$ for k$\geq$2
- $\mathrm{E}(Q) = k$
- $\mathrm{Var}(Q) = 2k$



- As usual, "converges" toward a normal distribution when k grows large.

Outline

1. The χ^2 distribution
2. Applications to statistical hypothesis test
   - Biased Coin
   - Adequation
   - Independence
   - Limitations
3. Other application of the χ^2 distribution
   - Student's law

A biased coin

Let's assume we are given a series of n coin toss. How could we check whether the coin is biased or not ?

A biased coin

Let's assume we are given a series of n coin toss. How could we check whether the coin is biased or not ?

The sample frequency of "Head" should be close to p when n is large.

- $H/n \sim \mathcal{N}\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$

A biased coin

Let's assume we are given a series of n coin toss. How could we check whether the coin is biased or not ?

The sample frequency of "Head" should be close to p when n is large.

- $H/n \sim \mathcal{N}\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$
- $\boxed{\mathcal{H}_0 : p = 1/2}$ then $\mathrm{P}\left(\left|\frac{H}{n} - 1/2\right| \leq \frac{1}{\sqrt{n}}\right) = 95\%.$

$\rightsquigarrow$ $\boxed{\text{Reject if } \notin [0.4, 0.6]}$

```r
set.seed(44); N = 100;
X=sample(x=c(0,1), size = N, prob=c(0.45,0.55), replace=T)
X
sum(X==1)/N
```

```
  [1] 0 1 1 0 1 1 1 1 1 1 1 1 0 1 1 1 1 1 1 0 1 1 1 1 1 1 0 1 0 0
 [38] 1 0 1 1 1 0 1 1 1 1 1 1 1 1 1 1 0 0 0 0 1 0 1 1 1 0 1 0 0 1 0
 [75] 1 1 1 1 1 0 1 0 1 1 1 1 0 1 0 0 1 1 1 0 1 1 0 0 1 1
[1] 0.67
```

we would then correctly reject the $\mathcal{H}_0$ hypothesis!

A biased coin again

```r
set.seed(41); N = 100;
X=sample(x=c(0,1), size = N, prob=c(0.45,0.55), replace=T)
sum(X==1)/N
```

```
[1] 0.51
```

If $p \approx 1/2$ there is a good chance we do not detect the bias (Type II error). ☹

A biased coin again

```r
set.seed(41); N = 100;
X=sample(x=c(0,1), size = N, prob=c(0.45,0.55), replace=T)
sum(X==1)/N
```

```
[1] 0.51
```

If $p \approx 1/2$ there is a good chance we do not detect the bias (Type II error). ☹

```r
set.seed(44); N = 100;
X=sample(x=c(0,1), size = N, prob=c(0.5,0.5), replace=T)
sum(X==1)/N
```

```
[1] 0.61
```

We may also incorrectly reject the $\mathcal{H}_0$ (Type I error). ☹

Trying to reject $\mathcal{H}_0$

	$\mathcal{H}_0$ True	$\mathcal{H}_0$ False
Reject	Type I error *(False positive)*	Correct *(True positive)*
Fail to reject	Correct *(True negative)*	Type II error *(False negative)*

- We only know the rejection probability when $\mathcal{H}_0$ holds True.
- Whenever $\mathcal{H}_0$ is False, the distribution of H depends on $p \neq 1/2$, which is unknown!

How to extend the idea to a dice ?

How to extend the idea to a dice ?

We could estimate p_1, p_2, p_3, p_4, p_5, and p_6

How to extend the idea to a dice ?

We could estimate p_1, p_2, p_3, p_4, p_5, and p_6

- Wait! Did we estimate the frequency of tails earlier ? p_6 is probably not needed.
- Our estimates are all correlated with each others! How do we combine these estimations into a single test ?

Outline

Adequation

- Suppose we have n independant random observations (X_j) classified into k classes with respective number of observations N_1, N_2, ..., N_k.
- Let's assume we know the theoretical probabilities and want to test the corresponding hypothesis

$$\boxed{\mathcal{H}_0 : \forall j, \mathrm{P}(X_j = 1) = p_1, \ldots, \text{ and } \mathrm{P}(X_j = k) = p_k}$$

Adequation

- Suppose we have n independant random observations (X_j) classified into k classes with respective number of observations $N_1, N_2, \ldots, N_k$.
- Let's assume we know the theoretical probabilities and want to test the corresponding hypothesis

$$\boxed{\mathcal{H}_0 : \forall j, \mathrm{P}(X_j = 1) = p_1, \ldots, \text{ and } \mathrm{P}(X_j = k) = p_k}$$

- We have $\frac{N_i}{n} \approx p_i$. For large n, $\frac{N_i}{n} - p_i$ follows a normal distribution (CLT) centered on 0 and with a variance of $p_i(1 - p_i)/n$.
- Let' build on this idea:
  - $Var(N_i - np_i) = np_i(1 - p_i)$. Hence,
  - $Var\left((N_i - np_i)^2\right) = n^2 p_i^2 (1 - p_i)^2$.
  - Therefore $\frac{(N_i - np_i)^2}{np_i} \sim \left(\mathcal{N}(0, (1 - p_i))\right)^2$.

Adequation

- Suppose we have n independant random observations (X_j) classified into k classes with respective number of observations $N_1, N_2, \ldots, N_k$.
- Let's assume we know the theoretical probabilities and want to test the corresponding hypothesis

$$\boxed{\mathcal{H}_0 : \forall j, \mathrm{P}(X_j = 1) = p_1, \ldots, \text{ and } \mathrm{P}(X_j = k) = p_k}$$

- We have $\frac{N_i}{n} \approx p_i$. For large n, $\frac{N_i}{n} - p_i$ follows a normal distribution (CLT) centered on 0 and with a variance of $p_i(1 - p_i)/n$.
- Let' build on this idea:
  - $Var(N_i - np_i) = np_i(1 - p_i)$. Hence,
  - $Var\left((N_i - np_i)^2\right) = n^2 p_i^2 (1 - p_i)^2$.
  - Therefore $\frac{(N_i - np_i)^2}{np_i} \sim \left(\mathcal{N}(0, (1 - p_i))\right)^2$.
- $T = \sum_{i=1}^{k} \frac{(N_i - np_i)^2}{np_i} \sim \chi_k^2$

Adequation

- Suppose we have n independant random observations (X_j) classified into k classes with respective number of observations $N_1, N_2, \ldots, N_k$.
- Let's assume we know the theoretical probabilities and want to test the corresponding hypothesis

$$\boxed{\mathcal{H}_0 : \forall j, \mathrm{P}(X_j = 1) = p_1, \ldots, \text{ and } \mathrm{P}(X_j = k) = p_k}$$

- We have $\frac{N_i}{n} \approx p_i$. For large n, $\frac{N_i}{n} - p_i$ follows a normal distribution (CLT) centered on 0 and with a variance of $p_i(1 - p_i)/n$.
- Let' build on this idea:
  - $Var(N_i - np_i) = np_i(1 - p_i)$. Hence,
  - $Var\left((N_i - np_i)^2\right) = n^2 p_i^2 (1 - p_i)^2$.
  - Therefore $\frac{(N_i - np_i)^2}{np_i} \sim \left(\mathcal{N}(0, (1 - p_i))\right)^2$.
- $T = \displaystyle\sum_{i=1}^{k} \frac{(N_i - np_i)^2}{np_i} \sim \chi^2_{k-1}$ (the last *correlated* term compensates for the others)

The χ^2 test

- Assume we know the theoretical frequencies p_i
- Count the number of occurences of each category
- Compute $T = \sum_{i=1}^{k} \frac{(N_i - np_i)^2}{np_i}$
- If all the $X_j \sim p$, then $T \sim \chi^2_k$ and $P(T < v) = 95\%$, with $v =$ qchisq(p=.95,df=k)

```
qchisq(p=.95,df=1)
qchisq(p=.95,df=3)
qchisq(p=.95,df=5)
```

```
[1] 3.841459
[1] 7.814728
[1] 11.0705
```

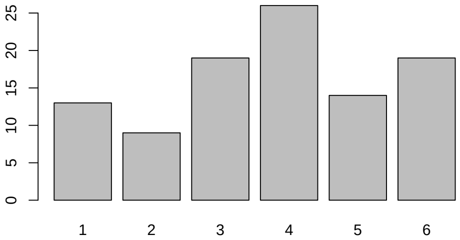
For an unbiased dice, it is "unlikely" that $T > 11.07$. If so reject the $\boxed{\mathcal{H}_0 : \text{unbiased}}$ hypothesis.

A biased dice

```
set.seed(44); N = 100;
X=sample(x=1:6, size = N, prob=c(.16,.16,.16,.16,.16,.2), replace=T
chisq.test(table(X),p=rep(1/6,times=6))
```

```
        Chi-squared test for given probabilities

data:  table(X)
X-squared = 10.64, df = 5, p-value = 0.059
```



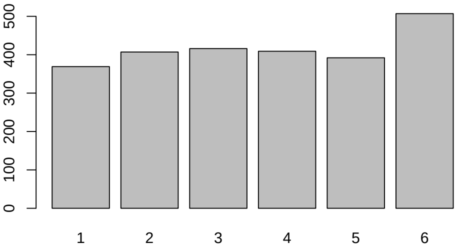
You cannot reject the hypothesis. And given the samples and your prior knowledge on the $\overline{\mathcal{H}_0}$, it's probably a good thing. ☺

A biased dice

```
set.seed(44); N = 2500;
X=sample(x=1:6, size = N, prob=c(.16,.16,.16,.16,.16,.2), replace=T)
chisq.test(table(X),p=rep(1/6,times=6))
```

```
        Chi-squared test for given probabilities

data:  table(X)
X-squared = 26.864, df = 5, p-value = 6.063e-05
```



26.8! The probability to get such a high value (or higher) is 0.00006. I believe this dice is biased.

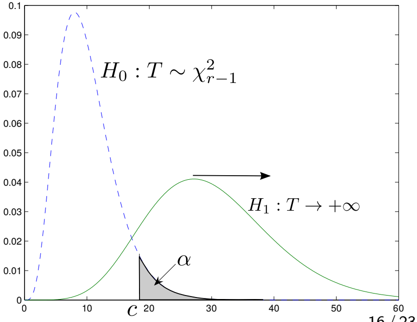
Testing through Goodness of Fit

Testing value T:

- What happens when $\mathcal{H}_0$ holds true? $T \sim \chi^2_{k-1}$
- What happens when $\mathcal{H}_0$ is false (e.g., $\pi_l \neq p_l$) ?

$$E(T) = \sum_{i=1}^{k} E\left(\frac{(N_i - np_i)^2}{np_i}\right) \geq E\left(\frac{(N_l - np_l)^2}{np_l}\right)$$

  - We have $E(N_l) = n\pi_l$ and $Var(N_l) = n\pi_l(1-\pi_l)$
  - $$E\left((N_l - np_l)^2\right) = \text{Var}(N_l - np_l) + E(N_l - np_l)^2$$
    $$= n\pi_l(1-\pi_l) + (n(\pi_l - p_l))^2$$
  - Therefore $\boxed{E(T) \geq n\frac{(\pi_l - p_l)^2}{p_l}}$



Outline

1. The χ^2 distribution
2. Applications to statistical hypothesis test
   - Biased Coin
   - Adequation
   - Independence
   - Limitations
3. Other application of the χ^2 distribution
   - Student's law

Setup

We measure $X_j \in \{A, B, C, D\}$ and $Y_j \in \{W, B, N\}$ and would like to know whether they are independent ($\mathcal{H}_0$) or not.

	A	B	C	D	total
White collar	90	60	104	95	349
Blue collar	30	50	51	20	151
No collar	30	40	45	35	150
Total	150	150	200	150	650

Problem:

- We do not know the p, (i.e., $\mathrm{P}(Y_j = W), \dots$)
  If we assume independance, let's use the sample frequency instead.
- Many of the cells are correlated.

$N_{A,W} = 90$ but it "should have been" $E_{A,W} = 150 \times \frac{349}{650} \approx 80.53$.

Therefore $\boxed{T = \sum_{c \in \{A,B,C,D\} \times \{W,B,N\}} \frac{(N_c - E_c)^2}{E_c} \sim \chi^2_6}$

χ^2 Independance Test

```
workers
chisq.test(workers)
```

```
                A  B   C  D
White collar 90 60 104 95
Blue collar  30 50  51 20
No collar    30 40  45 35

        Pearson's Chi-squared test

data:  workers
X-squared = 24.571, df = 6, p-value = 0.0004098
```

The probability of getting such a high value (or higher) for T is 0.0004098. This is unlikely, hence I decide to reject the independance hypothesis.

Outline

Limitation

- Random samples...
- Enough samples for the CLT to hold
  - More than 50 in total and more than 5 in each category ?
- Enough samples to discriminate from a close alternative
- Discrete values and not too many categories (remember how χ^2_k flattens with k)
- The probabilities (p_i) should be as close as possible to each others (rare categories will not help discrimination)
- Not too much samples...
  - If $n = 1,000,000$, the slightest difference will be overemphasized and it is likely that your samples will never match what you expected (your $\mathcal{H}_0$).

Outline

Student's law

The CLT allows to compute a confidence interval on an estimation of the expectation.

- It is centered on the sample mean
- The width is proportional to the standard deviation divided by the square root of the number of samples

Student's law

The CLT allows to compute a confidence interval on an estimation of the expectation.

- It is centered on the sample mean
- The width is proportional to the standard deviation divided by the square root of the number of samples
- How do we know the standard deviation ?
  - We can use the sample standard deviation but we have no idea of its distribution
  - Unless we assume X is normal, in which case
- If $S \sim \mathcal{N}$ and $Y \sim \chi^2_n$, then $\frac{S}{\sqrt{Y/n}} \sim$ t-Student.

This allows to account for the variance uncertainty.