# Computational Statistics

## Lecture 6: Two distributions, are they of the same kind ?

Raymond Bisdorff

University of Luxembourg

29 novembre 2019





## Content of Lecture 6







## Comparing statistical distributions

- Given two sequences of random numbers, we can ask the question : "*Are the two sequences drawn from a same random number generator, or from different generators ?*"
- In proper statistical terms : "*Can we disprove, to a certain required level of significance that two data sets are drawn from the same population distribution function ?*"
- Disproving the null hypothesis proves that the data are from different random distributions.
- Failing to disprove, on the othe hand, only shows that the data sets appear to be consistent with being generated from a same distribution function.





## Methodological approach

Four problems may appear from two dichotomies :

1. The data are either :
   1.1 continuous, or
   1.2 binned.
2. We wish to compare either
   2.1 one data set to a known distribution, or
   2.2 two equally unknown data sets.

## Statistical tests

- The usual test for differences between binned data is the Chi-square *goodness-of-fit* test.
- For continuous data as a function of a single variable, the usual test is the Kolmogorov-Smirnov test.
- One can always turn continuous data into binned data, by grouping the observed data into specified ranges of the continuous variable(s).
- There is however often some arbitrariness as how the bins should be chosen ; how many bins, with equal sizes or not ?
- Furthermore, binning always involves some loss of information. Even more, when uniform distributions of observations are not verified within all bins.
- Mind that statistical summaries are not truthful per se. They are merely numerical or graphical arguments supporting one or the other hypothesis concerning the observed data.











## Chi-square test against a known distribution

- Consider a random sequence grouped into $\upsilon$ bins.
- Suppose that $N_i$ is the number of events observed in the $i$th bin, and that $n_i$ is the number of expected events according to some known distribution. Note that the $N_i$'s are integers, while the $n_i$'s may not be.
- Then the Chi-square "*goodness-of-fit*" test statistic is :

$$\boxed{\chi^2 = \sum_{i=1}^{\upsilon} \frac{(N_i - n_i)^2}{n_i}}$$

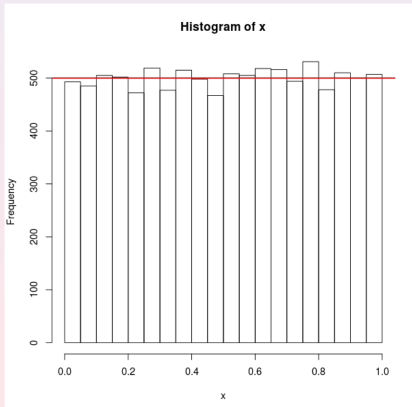
  where the sum runs over all $\upsilon$ bins.
- A value of $\chi^2 \gg \upsilon$ indicates that a "*goodness-of-fit*" is rather unlikely.



## Uniformity Chi-Square *goodness-of-fit* Test in R

Let us test if the R `runif` generator is giving consistent data with a uniform distribution. The R `chisq.test` method implements this *goodness-of-fit* test.

```
> nSim = 10^4
> x = runif(nSim)
> freq = hist(x)
> Ni = freq$counts
> upsilon = length(Ni)
  [1] 20
> ni = rep(nSim/upsilon,upsilon)
> chi2 = sum((Ni-ni)^2/ni)
  [1] 18.4988
> df = upsilon - 1
> pvalue = 1.0 - pchisq(chi2,df)
  [1] 0.4893842
> chisq.test(Ni)
  X-squared = 18.4988 df = 19
  p-value = 0.4893842
```



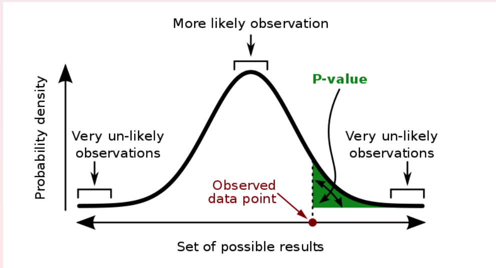
## Chi-square Test – continue

- Any term $i$ with $0 = n_i = N_i$ should be omitted from the sum.
- A term with $n_i = 0$ and $N_i \neq 0$ gives an infinite $\chi^2$, as it should, since in this case the $N_i$'s cannot possibly be drawn from these $n_i$'s.
- The $P(\chi^2|\upsilon)$ probability function with degree of freedom $\upsilon$ is the probability that the sum of the squares of $\upsilon$ standard Gaussian variables of unit variance and 0 mean will be greater than $\chi^2$.
- The terms in the sum of the $\chi^2$ measure are only good approximations of squares of random standard normal variables when $N_i \gg 1$ in each bin.
- Usually, the binning process gives a constrained last bin content. Hence, the degree of freedom of $P(\chi^2|\upsilon)$ is only $\upsilon - 1$ !



## Significance of the *goodness-of-fit* test

- The $P(\chi^2|\upsilon)$ probability function gives via the p-value a good estimate for the actual significance of the chi-square goodness-of-fit test.
- The p-value equals the probability that the Chi-square test may give, under the "*goodness-of-fit*" hypothesis, a result greater or equal than $x$ : $\mathcal{P}(\chi^2|\upsilon \geq x) \ = \ 1.0 - \mathcal{P}(\chi^2|\upsilon \leq x)$.
- The higher, resp. the smaller, the $p$-value, the more the goodness-of-fit is likely, resp. unlikely.
- If a certain significance level is required, like 95% for instance, then the *goodness-of-fit* hypothesis is rejected if the $p$-value is smaller than 5%.




Source : https ://en.wikipedia.org/wiki/P-value



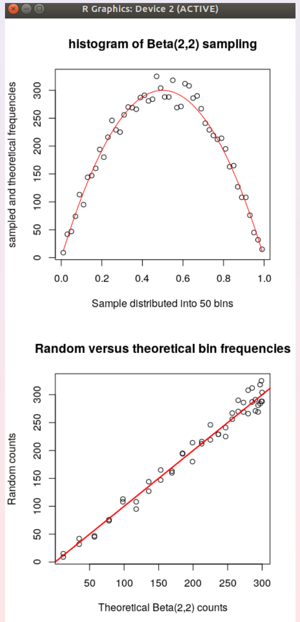
### Exercise (Chi-square "*goodness-of-fit*" tests)

1. *How to apply a Chi-square "goodness-of-fit" tests to samples taken with a* $\mathcal{B}(2,2)$ *random number generator ?*
2. *How to check the accuracy of random sampling from the empirical random law shown on slide 12/34 of lecture 3 ?*
3. *May the random sequences obtained with a Mersenne twister RNG versus the ones obtained from a linear congruational RNG be discriminated by the Chi-square "goodness-of-fit" test ?*
4. *What is the distribution of p-values for samples of size* $n = 10^4$ *of uniform random numbers generated with* `runif(n)` *?*



## Checking *goodness-of-fit* of a $\mathcal{B}(2,2)$ sample

```
> par(mfrow=c(2,1))
> nSim = 10^4
> xb = rbeta(nSim,2,2)
> h = hist(xb,breaks=50,plot=F)
> plot(h$mids,h$counts)
> thcounts =
+   dbeta(h$mids,2,2)*0.02*nSim
> lines(h$mids,thcounts,col="red")
> plot(thcounts,h$counts)
> abline(0,1,col="red",lwd=2)
> Ki2=sum((h$counts-thcounts)^2/thcounts)
[1] 46.494
> pval = 1-pchisq(Ki2,length(h$counts)-1)
[1] 0.5753
> chisq.test(h$counts,
+ p=dbeta(h$mids,2,2)*0.02,rescale.p=T)
X-squared = 46.503, df = 49,
p-value = 0.5749
```



## Comparing *two* binned data sets of same size

- Let $R_i$ be the number of events observed in the $i$th bin for the first data set, and let $S_i$ be the number of events in the same bin for data set two.
- Then the chi-square "*goodness-of-fit*" test statistic is :

$$\boxed{\chi^2 = \sum_{i=1}^{\upsilon} \frac{(R_i - S_i)^2}{R_i + S_i}}$$

  where the sum runs over all $\upsilon$ bins.
- If the data were collected in such a way that the sum of $R_i$'s is necessarily equal to the sum of the $S_i$'s, then the number of degrees of freedom is one less than the number $\upsilon$ of bins.





## Comparing *two* binned data sets of different size

- Let $R_i$ be the number of events observed in the $i$th bin for the first data set, and let $S_i$ be the number of events in the same bin for data set two.
- Then the chi-square "*goodness-of-fit*" test statistic is :

$$\boxed{\chi^2 = \sum_{i=1}^{\upsilon} \frac{(\sqrt{S/R}R_i - \sqrt{R/S}S_i)^2}{R_i + S_i}}$$

  where $R := \sum_i R_i$ and $S := \sum_i S_i$.
- The number of degrees of freedom is still one less than the number $\upsilon$ of bins.





## Problem with small number of counts

- When significant fractions of bins have a small number of counts ($\leqslant 10$, say), then $\chi^2$ statistics are not well approximated by a chi-square probability function.
- Under the "*goodness-of-fit*" hypothesis, the count in an individual bin, $N_i$, is following a Poisson law with $\lambda = n_i$ and each term $(N_i - n_i)^2/n_i$ has $\mu = 1$ and $\sigma^2 = 2 + 1/n_i$.
- Each term in the $\chi^2$ statistic adds, on average, 1 to its value, and slightly more than 2 to its variance.
- But, the variance of the chi-square probability function is exactly twice its mean. If a significant fraction of $n_i$'s are small, then quite probable values of the $\chi^2$ statistic will appear to lie farther out on the tail than they actually are.
- Thus, the "*goodness-of-fit*" hypothesis may be rejected even when it is true.





## Remedies with small number of counts

- Regroup the bins with small number of counts.
- When $\upsilon$, the number of bins, is large ($> 30$), the central limit theorem implies that the $\chi^2$ statistic gets approximately a Gaussian distribution :

$$\chi^2 \rightsquigarrow \mathcal{N}\Big(\upsilon, \big[2\upsilon + \sum_i n_i^{-1}\big]^{1/2}\Big),$$

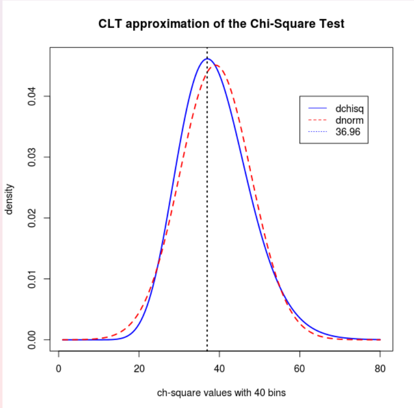
  and $p$-values may be computed as a complement of the corresponding cumulated Gaussian distribution function.
- In the case of *two* binned data sets :

$$\sum_i n_i^{-1} \to \Big[\frac{(R-S)^2}{RS} - 6\Big] \sum_i \frac{1}{R_i + S_i}$$

## Remedies for small number of counts in R

A $P(\chi^2|\upsilon)$ cdf may be approximated with a Gaussian cdf when $\upsilon > 30$ as shown in R plot below.

```
> breaks = seq(0,1,0.025)
> freq = hist(x,breaks)
> Ni = freq$counts
> upsilon = length(freq$breaks)-1
> ni = rep(Nsim/upsilon,upsilon)
> chi2 = sum((Ni-ni)^2/ni)
  [1] 36.96
> df = upsilon - 1
> pvalue = 1.0 - pchisq(chi2,df)
  [1] 0.5632565
> sigma = sqrt(2*df+sum(1/ni))
> npvalue = 1.0 -
+        pnorm(chi2,df,sigma)
  [1] 0.5912447
```








## RNG Quality : Testing equidistribution

Let $\langle U_n \rangle = [u_0, u_1, u_2, ...]$ be a sequence of random numbers from the float interval $[0.0; 1.0)$ apparently generated in a uniformly manner.
To test the quality of the random generator, we consider the auxiliary sequence $\langle Y_n \rangle = [y_0, y_1, y_2, ...]$ defined by the rule $y_n = \lfloor d \times u_n \rfloor$, where $d$ is a positive integer – usually 64, 100, or 128 – also called the *discrete grain* of the generator.
When sequence $\langle U_n \rangle$ is indeed uniformly distributed, we will observe a sequence $\langle Y_n \rangle$ of equidistributed integers between 0 and $d-1$.
The quality of a given random generator may now be assessed with a two-tailed Chi-square "*goodness-of-fit*" test between the empirical $N_i$ distribution and the theoretical uniform $n_i = 1/d$ distribution.

A $p$ -value below 5% or above 95% indicates the very likeliness of a suspicious non-randomness in $\langle U_n \rangle$.





## RNG Quality : Serial test

- We reconsider the auxilliary $\langle Y_n \rangle$ sequence with discrete grain $d$ and count the number of times the pair $(y_{2j}, y_{2j+1}) = (q, r)$ occurs, for $0 \leq j < n/2$, $q \neq r$ and $0 \leq q, r \leq d$.
- These counts are to be made for each pair of integers $(q, r)$ with $0 \leq q, r \leq d$, and the Chi-square "*goodness-of-fit*" test is applied to these $k = d^2$ categories with theoretical uniform relative frequency $1/d^2$ in each category.
- To keep the length $n$ of the random sequence large compared to $k$, $d$ will be chosen of smaller value than for the equidistributional test.





## RNG Quality : Gap test

- Another test is to examine the length of "gaps" between occurences of $u_j$ in a certain range. If $\alpha$ and $\beta$ are two real numbers with $0 \leq \alpha < \beta \leq 1$, we want to consider the lengths of consecutive subsequences $[u_j, u_{j+1}, ..., u_{j+r}]$ in which the consecutive $r$ values $u_{j+k}$, for $k = 1, ...r$, remain between $\alpha$ and $\beta$. This situation will be counted as a gap of length $r$.
- With given values $\alpha$ and $\beta$ and a maximal gap length $t$, let $C_r$ for $r = 0, ..., t-1$ count the occurences of gaps of length $0, ..., t-1$, and $C_t$ the gaps of length $r \geq t$. If $p = \beta - \alpha$, the theoretical counts for each gap length $r$, is $p_r = p(1-p)^r$ for $0 \leq r < t-1$ and $p_t = (1-p)^t$.
- Again, a Chi-square "*goodness-of-fit*" test, comparing the $C_r$ with the $p_r$ distribution may be used in order to assess the likeliness of a suspicious non-randomness of the gap lengths observed in the sequence $\langle U_n \rangle$.

## RNG Quality : Coupon collector's test

- This test relates the frequency test to the previous gap test. We use the auxiliary sequence $\langle Y_n \rangle$ and we observe the lengths of subsequences $y_{j+1}, y_{j+2}, \ldots, y_{j+r}$ that are required to get a complete set of integers – a coupon collector seqment – from 0 to $d-1$.
- With a given maximal subsequence length $t$, let $C_r$ for $r = d, \ldots, t-1$ count the occurences of coupon collector segments of length $d, d+1, \ldots, t-1$, and $C_t$ the segments of length $r \geq t$.
- The theoretical count for each coupon collector segment of length $r$, is

$$p_r = \frac{d!}{d^r}\left\{\begin{matrix} r-1 \\ d-1 \end{matrix}\right\}, \quad d \leq r < t-1; \quad p_t = 1 - \frac{d!}{d^r}\left\{\begin{matrix} r \\ d \end{matrix}\right\}.$$

- Similarly, a Chi-square "*goodness-of-fit*" test, comparing the empirical $C_r$ with the theoretical $p_r$ distribution, may be used in order to assess the likeliness of a suspicious non-randomness of the coupon collector segments.





## RNG Quality : Up and down runs test

- A sequence $\langle U_n \rangle$ of uniform random numbers may also be tested for "*runs up*" and "*runs down*" segments, by examining the length of monotone portions of it. Let $[u_{j+0}, u_{j+1}, \ldots, u_{j+r}]$ be a subsequence of length $r$ such that either $u_{j+0} \geq u_{j+1} \geq \ldots \geq u_{j+r}$, or, $u_{j+0} \leq u_{j+1} \leq \ldots \leq u_{j+r}$.
- Given a maximal subsequence length $t$, let $C_r$ for $r = 1, \ldots, t-1$ count the occurences of separated monotone, either up, or, down runs of length $1, 2, \ldots, t-1$, and $C_t$ the same runs of length $r \geq t$.
- Assuming that a monotone run of length $r$ occurs with probability $1/r! - 1/(r+1)!$, the theoretical relative count for each length $r$, gives $p_r = 1/r! - 1/(r+1)!$ for $r < t$ and $p_t = 1/t!$.
- And, again, we may use a Chi-square "*goodness-of-fit*" test, comparing the empirical $C_r$ with the theoretical $p_r$ distribution, for assessing the likeliness of a suspicious non-randomness of "*runs up*" or "*runs down*" segments.











## Kolmogorov-Smirnov Test

- The ordered list of data points is converted into a cumulative distribution function of the probability distribution from which it has been drawn.
- If the $N$ events are located at points $x_i$, $i = 1, \ldots, N$, then $S_N(x)$ is giving the fraction of points to the left of a given value $x$.
- The Kolmogorov-Smirnov statistic $D$ is defined as the maximum value of the absolute difference between two cumulative distribution functions.
- When comparing $S_N(x)$ to a known cdf $P(x)$, the K-S statistic is

$$\boxed{D = \max_{-\infty < x < +\infty} |S_N(x) - P(x)|}$$

- For comparing two different cdf's, the K-S statistic is

$$\boxed{D = \max_{-\infty < x < +\infty} |S_{N_1}(x) - S_{N_2}(x)|}$$

# Kolmogorov-Smirnov Test – continue

- Testing the p-value significance of the K-S test is done with the complement $Q_{KS}(z) = 1 - P_{KS}(z)$ of the cdf $P_{KS}(z)$ of the K-S distribution for $z > 0$ :

$$P_{KS}(z) \ = \ 1 - 2\sum_{j=1}^{\infty}(-1)^{j-1}exp(-2j^2z^2)$$

- The K-S statistic is invariant under reparametrization of the data set points. $D$ remains the same when locally stretching and sliding the $x$ axis. Using for instance $x$ or $\log x$ in $D$ will result in the same significance of the test.





# Kolmogorov-Smirnov Test in R

- The $D$ observed and its $p$-value as disproof of the null hypothesis that the distributions under review are the same is given by the R `ks.test` procedure.

```
> x = rnorm(50)
> ks.test(x,"pnorm")
  D = 0.101, p-value = 0.6498
> y = runif(30,-2.5,2.5)
> plot(ecdf(x),col="blue")
> plot(ecdf(y),add=T,col="red")
> ks.test(x,y,exact=T)
  D = 0.3267, p-value = 0.02926
```