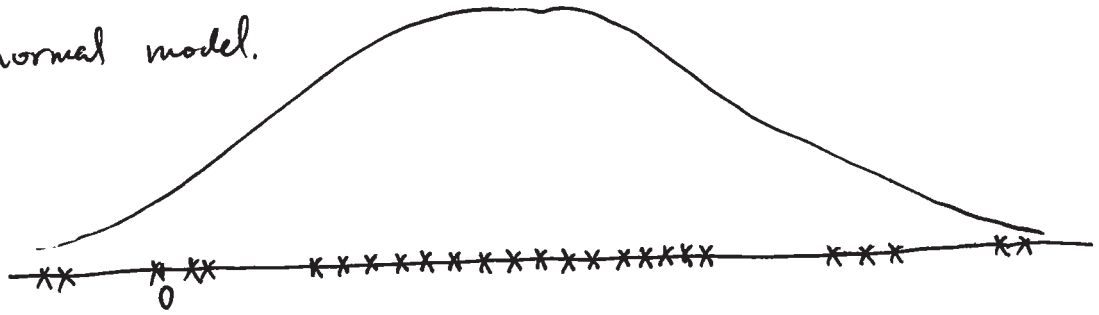
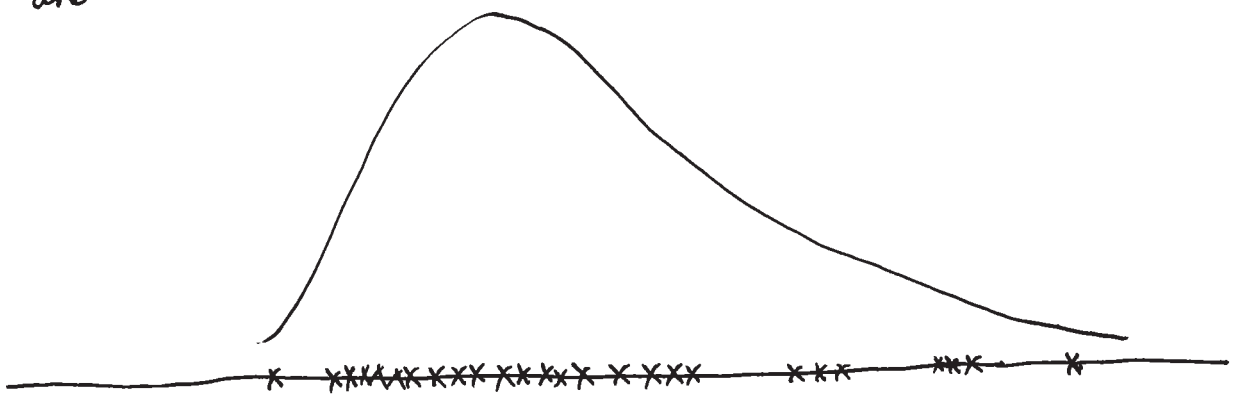
model checking:

In many examples. it is convenient to fit Normal model, i.e. samples follow a normal model. However, the normal assumption may not be always correct. For example, if we observe the following samples. we should ask if they really follow a normal model.



Clearly, the above samples are possible to be normal since they stay reasonably under a bell shaped curve.

But, are the samples below normal?



they may not since the curve is right skewed. (not symmetric)

Question: given a set of samples, are they normal?

this is the socalled model checking problem.

Suppose a sample $X_1, X_2, \cdots, X_n$ is given to you,

we have the following two choices:

they are normal	they are not normal

if the samples $X_1 \cdots X_n$ are normal from $N(\mu, \sigma_0^2)$.

then. $\frac{X_1-\mu}{\sigma_0} \sim N(0,1)$.

$\frac{X_2-\mu}{\sigma_0} \sim N(0,1)$.

$\vdots$

$\frac{X_n-\mu}{\sigma_0} \sim N(0,1)$

So $\left(\frac{X_1-\mu}{\sigma_0}\right)^2 + \left(\frac{X_2-\mu}{\sigma_0}\right)^2 + \cdots + \left(\frac{X_n-\mu}{\sigma_0}\right)^2 \sim \chi^2_{(n)}$

the χ^2 - distribution of $df = n$.

The left hand side cannot be directly used to check normality since the unknown μ is there. To fix this unknown μ. we have to cheat. the smart way to cheat is to replace μ by its MLE $\bar{X}$. So. we have

$$\left(\frac{X_1-\bar{X}}{\sigma_0}\right)^2 + \left(\frac{X_2-\bar{X}}{\sigma_0}\right)^2 + \cdots + \left(\frac{X_n-\bar{X}}{\sigma_0}\right)^2 \sim \chi^2(n-1)$$

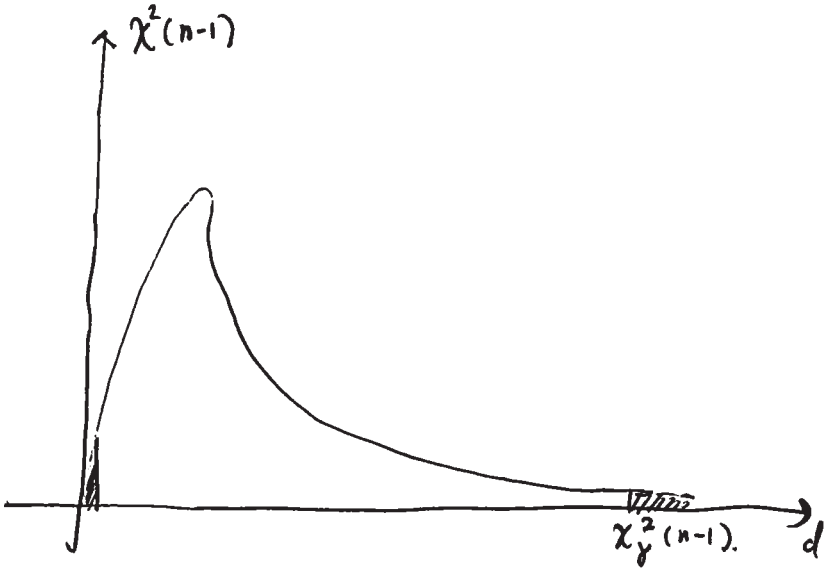
the quantity on the right is denoted as

$$D = \left(\frac{X_1 - \bar{X}}{\sigma_0}\right)^2 + \left(\frac{X_2 - \bar{X}}{\sigma_0}\right)^2 + \cdots + \left(\frac{X_n - \bar{X}}{\sigma_0}\right)^2.$$

So under the normality assumption on the samples.

$$D \sim \chi^2(n-1).$$

~~If~~ the quantity D is always ~~be~~ found in practice. since it is ~~only~~ only depending on the samples. ~~If~~



If we observe D in the main body of the $\chi^2(n-1)$. -curve. we believe the normality assumption is reasonable.

If we observe D in the extreme regions (like the shaded ones in the above), then we belive some strange or extreme has happened. so that the normality hypothesis is rejected.

We usually compare D with the ~~γ~~ γ cut-off. $\chi^2_{0.95}(n-1)$. ~~or the~~

In practice. $\gamma = 0.95$ is preferred.

STAT 417. Lecture Note 22.

Eg1. If we observe a set of samples with $\sigma_0^2 = 1$.

1.58, 0.98, 0.45, −0.40, −0.16 (so $\bar{x} = 0.49$).

then. $D = (1.58 - 0.49)^2 + (0.98 - 0.49)^2 + (0.45 - 0.49)^2 + (-0.40 - 0.49)^2$

$+(-0.16 - 0.49)^2 = 2.65 < \chi^2_{0.95}(4) = 9.49.$

So the samples seem to be follow normal model.

Eg2. If we have $n = 10$ samples with $\sigma_0^2 = 4$.

Suppose the sample variance $S^2 = 10$

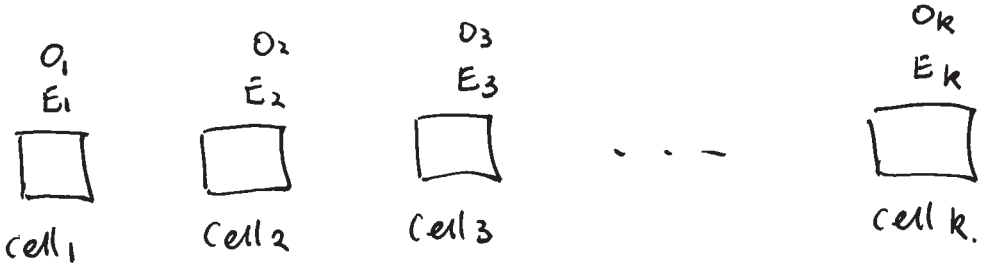
Are the samples from normal model?

Sol: $D = \frac{1}{\sigma_0^2}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{(n-1)S^2}{\sigma_0^2} = \frac{(10-1)\times 10}{4} = 22.5 > \chi^2_{0.95}(9) = 16.92.$

So the samples may not follow normal model.

χ^2 goodness of fit test (A review of STAT 350).

n elements in k cells.



O_i = observed frequency in cell i.

$i = 1, 2, \ldots, k$.

E_i = Expected frequency in cell i.

Then the χ^2 goodness of fit test is

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \approx \chi^2(k-1).$$

multinomial model checking.

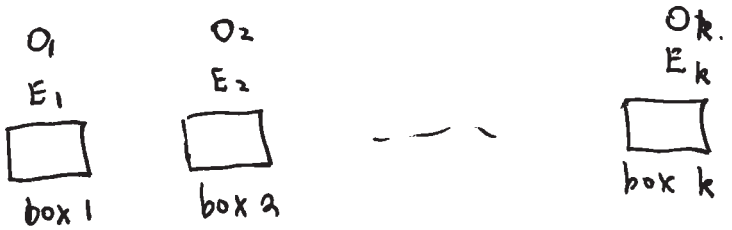
Suppose k boxes. Let O_i = # of balls in the ith box.

$i = 1, 2, \ldots, k$. $\quad n = O_1 + O_2 + \cdots + O_k$.

Question: Is $(O_1, O_2, \ldots O_k)$ from multinomial$(n; p_1, p_2, \ldots p_k)$?

where $p_1, \ldots, p_k > 0$ are known. $p_1 + p_2 + \cdots + p_k = 1$.

under $(O_1, \ldots, O_k) \sim$ multinomial $(n; p_1, p_2, \ldots, p_k)$.



$E_1 = np_1, \quad E_2 = np_2, \ldots, E_k = np_k$.

so. the χ^2 goodness of fit test is

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - np_i)^2}{np_i}$$

The quantity $R_i = \dfrac{O_i - np_i}{\sqrt{np_i(1-p_i)}}$. is the standized residual.

Eg3. If $(O_1, O_2, O_3) = (45, 40, 15)$.

check $(O_1, O_2, O_3) \sim$ multinomial $(100; 0.6, 0.3, 0.1)$

sol: $\chi^2 = \frac{(45-60)^2}{60} + \frac{(40-30)^2}{30} + \frac{(15-10)^2}{10} = 4.58 > k^2_{0.95}(2) = 5.99$.

So the samples do not follow the multinomial model.

the standized residual is

$$R_1 = \frac{45-60}{\sqrt{100(0.6)(0.4)}} = -3.06.$$

$$R_2 = \frac{40-30}{\sqrt{100(0.3)(0.7)}} = 2.18.$$

$$R_3 = \frac{15-10}{\sqrt{100(0.1)(0.9)}} = 1.67.$$

Exponential model checking.

Eg 4.

Suppose we have $X_1, X_2, \cdots, X_{100}$ i.i.d samples.

Question: do they follow Exponential(1) ?

Idea: use χ^2-test.

First, construct cells $B_1 = [0,1]$, $B_2 = (0,2]$, $B_3 = (2,5]$, $B_4 = (5,\infty)$.

~~define~~ let O_i = # of samples in the ith box, $i = 1,2,3,4$.

Record, filed E_1, E_2, E_3, E_4.

E_i = expected # of samples in B_i. $i = 1,2,3,4$.

$$p_1 = \int_0^1 e^{-x}\,dx = 1 - e^{-1} = 0.632. \quad (P(X_1 \in B_1)).$$

$$p_2 = P(X_2 \in B_2) = \int_1^2 e^{-x}\,dx = e^{-1} - e^{-2} = 0.232.$$

$$p_3 = P(X_3 \in B_3) = \int_2^5 e^{-x}\,dx = e^{-2} - e^{-5} = 0.129$$

$$p_4 = P(X_4 \in B_4) = \int_5^\infty e^{-x}\,dx = e^{-5} = 0.007.$$

So $E_1 = (100)(0.632) = 63.2$. $E_2 = (100)(0.232) = 23.2$.

$E_3 = (100)(0.129) = 12.9$, $E_4 = (100)(0.007) = 0.7$.

If we observed that $(O_1, O_2, O_3, O_4) = (60, 25, 14, 1)$.

Then. $\chi^2 = \frac{(60-63.2)^2}{63.2} + \frac{(25-23.2)^2}{23.2} + \frac{(14-12.9)^2}{12.9} + \frac{(1-0.7)^2}{0.7}$

$= 0.524 < \chi^2_{0.95}(3) = 7.81.$

So. the samples may come from Exponential(1).

$$R_1 = \frac{60-63.2}{\sqrt{100(0.632)(1-0.632)}} = -0.664.$$

$$R_2 = \frac{25-23.2}{\sqrt{100(0.232)(1-0.232)}} = 0.426.$$

$$R_3 = \frac{14-12.9}{\sqrt{100(0.129)(1-0.129)}} = 0.328.$$

$$R_4 = \frac{1-0.7}{\sqrt{100(0.007)(1-0.007)}} = 0.360.$$

Ex1. If Joe has $n=100$ kittens, among who 40 are female.

Let $X=$ # of female kittens. Test $X \sim \text{Binomial}(100, 0.5)$.

Sol:

$O_1 = 40$ $O_2 = 60$

$E_1 = 50$ $E_2 = 50$.

So. $\chi^2 = \frac{(40-50)^2}{50} + \frac{(60-50)^2}{50} = 4 > \chi^2_{0.95}(1) = 3.84$.

$R_1 = \frac{40-50}{\sqrt{100(0.5)(0.5)}} = -2.$

$R_2 = \frac{60-50}{\sqrt{100(0.5)(0.5)}} = 2.$