4.3.2 Two-sample t-test assuming equal variances (ttest2 in Matlab)

Remove the assumption that the samples are paired. So, we have two independent normal samples $(X_1, \dots, X_n)$ and $(Y_1, \dots, Y_m)$. We know:

$$A_X = \frac{\sqrt{n}(\hat{\mu}_X - \mu_X)}{\sigma_X} \sim N(0,1), \quad B_X = \frac{n\hat{\sigma}_X^2}{\sigma_X^2} \sim \chi^2_{n-1}$$

$$A_Y = \frac{\sqrt{m}(\hat{\mu}_Y - \mu_Y)}{\sigma_Y} \sim N(0,1), \quad B_Y = \frac{m\hat{\sigma}_Y^2}{\sigma_Y^2} \sim \chi^2_{m-1}$$

and A_X, A_Y, B_X, B_Y are independent.

So, $\dfrac{\hat{\mu}_X - \mu_X}{\sigma_X} \sim N\left(0, \frac{1}{n}\right)$, $\dfrac{\hat{\mu}_Y - \mu_Y}{\sigma_Y} \sim N\left(0, \frac{1}{m}\right) \Rightarrow$

$$\Rightarrow \frac{\hat{\mu}_X - \mu_X}{\sigma_X} - \frac{\hat{\mu}_Y - \mu_Y}{\sigma_Y} \sim N\left(0, \frac{1}{n} + \frac{1}{m}\right)$$

$$\Rightarrow A = \left(\frac{\hat{\mu}_X - \mu_X}{\sigma_X} - \frac{\hat{\mu}_Y - \mu_Y}{\sigma_Y}\right) \Big/ \left(\frac{1}{n} + \frac{1}{m}\right)^{1/2} \sim N(0,1)$$

Also, by definition of χ^2-distributions,

$$B = \frac{n\hat{\sigma}_X^2}{\sigma_X^2} + \frac{m\hat{\sigma}_Y^2}{\sigma_Y^2} \sim \chi^2_{m+n-2}$$

So, $\dfrac{A}{\sqrt{\frac{B}{n+m-2}}} \sim t_{n+m-2}$.

However, σ_x^2 and σ_y^2 are unknown, so we cannot ~~compute this expression.~~

Assume $\sigma_x^2 = \sigma_y^2 = \sigma^2$. Then

$$\frac{A}{\sqrt{\frac{B}{n+m-2}}} = \frac{\frac{\frac{(\hat{\mu}_x - \hat{\mu}_y) - (\mu_x - \mu_y)}{\cancel{\sigma}}}{\left(\frac{m+n}{mn}\right)^{1/2}}}{\sqrt{\frac{\frac{n\hat{\sigma}_x^2 + m\hat{\sigma}_y^2}{\cancel{\sigma^2}}}{n+m-2}}} =$$

$$= \underbrace{\left(\frac{mn(m+n-2)}{m+n}\right)^{1/2} \cdot \frac{(\hat{\mu}_x - \hat{\mu}_y) - (\mu_x - \mu_y)}{\left(n\hat{\sigma}_x^2 + m\hat{\sigma}_y^2\right)^{1/2}}}_{\text{depends only on } \mu_x - \mu_y} \sim t_{n+m-2}$$

So, we can test the hypotheses about this difference based on the t-statistic:

$$T = \left(\frac{mn(m+n-2)}{m+n}\right)^{1/2} \cdot \frac{\hat{\mu}_x - \hat{\mu}_y}{\left(n\hat{\sigma}_x^2 + m\hat{\sigma}_y^2\right)^{1/2}}$$

For example, if you want to test $H_0: \mu_x = \mu_y$ (i.e. $\mu_x - \mu_y = 0$), we use the t-test:

$$\delta = \begin{cases} H_0, & \text{if } |T| \le c \\ H_1, & \text{if } |T| > c \end{cases}$$

where c is found from $2t_{n+m-2}(c, \infty) = \alpha$ (just as in 4.1 ①, the only change is the ~~# of degrees~~ of freedom!)

<u>Note</u>: There is an approximation t-test which handles even the case when the variances are not assumed to be equal... too complicated... no time. "Satterthwaite approximation"

4.3.3 Comparing the variances of two normal distributions: F-test (vartest2 in Matl

Suppose we have two normal samples $(X_1, \dots, X_n)$ and $(Y_1, \dots, Y_m)$, and we want to test $H_0: \sigma_X = \sigma_Y$. or $H_0: \sigma_X \le \sigma_Y$.

We know: $B_X = \frac{n\hat{\sigma}_X^2}{\sigma_X^2} \sim \chi^2_{n-1}$, $B_Y = \frac{m\hat{\sigma}_Y^2}{\sigma_Y^2} \sim \chi^2_{m-1}$.

The ratio
$$\left(\frac{B_X}{n-1}\right) \Big/ \left(\frac{B_Y}{m-1}\right) = \frac{n(m-1)\hat{\sigma}_X^2}{m(n-1)\hat{\sigma}_Y^2} \cdot \frac{\sigma_Y^2}{\sigma_X^2} \sim F_{n-1,m-1}$$

is the F-distr. with $(n-1, m-1)$ degrees of freedom.

We use the following statistic:

$$F = \frac{n(m-1)}{m(n-1)} \frac{\hat{\sigma}_X^2}{\hat{\sigma}_Y^2} \sim \frac{\sigma_X^2}{\sigma_Y^2} F_{n-1,m-1}$$

When $\sigma_X^2 = \sigma_Y^2$ we have $F \sim F_{n-1,m-1}$

① $(H_0: \sigma_X = \sigma_Y)$ We use the following test:

$$\delta = \begin{cases} H_0, & \text{if } c_1 \le F \le c_2 \\ H_1, & \text{if } F < c_1 \text{ or } F > c_2. \end{cases}$$

Thresholds c_1, c_2 should satisfy the condition:

$$\alpha = \mathbb{P}(\delta = H_1 / H_0) = \mathbb{P}(F < c_1 \mid \sigma_X = \sigma_Y) + \mathbb{P}(F > c_2 \mid \sigma_X = \sigma_Y) =$$
$$= F_{n-1,m-1}(0, c_1) + F_{n-1,m-1}(c_2, \infty)$$

So, how do we find c_1 and c_2? Set $F_{n-1,m-1}(0, c_1) = F_{n-1,m-1}(c_2, \infty) = \frac{\alpha}{2}$

② $(H_0: \sigma_X \le \sigma_Y)$ We use the following test.

$$\delta = \begin{cases} H_0, & \text{if } F \le c \\ H_1, & \text{if } F > c \end{cases}$$

where c can be found from $\alpha = \mathbb{P}(\delta = H_1 | H_0) = \mathbb{P}(F > c \mid \sigma_X = \sigma_Y) = F_{n-1,m-1}(c, \infty)$

$\boxed{5}$ Testing simple hypothesis. Bayes decision rules

<u>Setup</u>: We have an i.i.d. sample $(X_1, \ldots, X_n) \in S$ from <u>unknown distribution</u> $\mathbb{P}$.
Suppose $\mathbb{P}$ belongs to a set of k specified distributions $\{\mathbb{P}_1, \ldots, \mathbb{P}_k\}$.
Then, given the sample $(X_1, \ldots, X_n)$ we have to decide among k hypotheses

$$\begin{cases} H_1 : \mathbb{P} = \mathbb{P}_1 \\ H_2 : \mathbb{P} = \mathbb{P}_2 \\ \quad \vdots \\ H_k : \mathbb{P} = \mathbb{P}_k \end{cases}$$

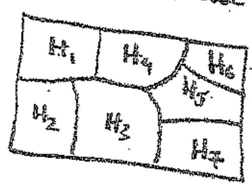
, i.e. we need a test $\delta : S \to \{H_1, \ldots, H_k\}$

<u>Def</u>: <u>Error of type i</u>, $i = \overline{1,k}$, is $\mathbb{P}(\delta \neq H_i | H_i) = \mathbb{P}_i(\delta \neq H_i) = \alpha_i$

<u>Def.</u> $(k=2)$ $\alpha_1 = \mathbb{P}_1(\delta \neq H_1)$ is also called the <u>size</u> or <u>level of significance</u> of test δ.

$\beta = 1 - \alpha_2 = 1 - \mathbb{P}_2(\delta \neq H_2) = \mathbb{P}_2(\delta = H_2)$ is also called the <u>power</u> of δ.

Ideally, we want to minimize errors of all types. This is usually impossible, so there is a trade-off.



$S = \{(x_1, \ldots, x_n)\}$

<u>Question</u>: How do we compare decision rules?

<u>Bayesian approach</u>: Consider k weights $\xi(i)$, $i = \overline{1,k}$ s.t. $\sum_{i=1}^{k} \xi(i) = 1$.
Think of $\xi(i)$ as a priori probability on the set of k hypothesis that represent their relative importance. Then the <u>Bayes error</u> is

$$\alpha(\xi) = \sum_{i=1}^{k} \xi(i) \alpha_i = \sum_{i=1}^{k} \xi(i) \mathbb{P}_i(\delta \neq H_i) \qquad \text{(simply a weighted error)}$$

<u>Def.</u> Test δ that minimizes $\alpha(\xi)$ is called a <u>Bayes decision rule</u>.

Theorem 1 Assume that each distribution $\mathbb{P}_i$ has a p.d.f. $f_i(x)$.

A test δ that chooses H_j when

$$\xi(j) f_j(X_1)\cdot\ldots\cdot f_j(X_n) = \max_{i=\overline{1,k}} \xi(i) f_i(X_1)\cdot\ldots\cdot f_i(X_n)$$

is the Bayes decision rule.

Note: If the max. is achieved simultaneously on several j's, pick any one at random.

proof: The Bayes error is $\alpha(\xi) = \sum_{i=1}^{k} \xi(i)\mathbb{P}_i(\delta \neq H_i) =$

$$= \sum_{i=1}^{k} \xi(i) \int \underbrace{I(\delta\neq H_i)}_{\text{indicator variable of the event } \delta\neq H_i} f_i(x_1) f_i(x_2)\ldots f_i(x_n)\, dx_1\ldots dx_n =$$

$$= \int \sum_{i=1}^{k} \xi(i)\, f_i(x_1)\cdot\ldots\cdot f_i(x_n) \left(1 - I(\delta = H_i)\right) dx_1\ldots dx_n =$$

$$= \int \sum_{i=1}^{k} \xi(i) f_i(x_1)\cdot\ldots\cdot f_i(x_n)\, dx_1\ldots dx_n - \int \sum_{i=1}^{k} \xi(i) f_i(x_1)\cdot\ldots\cdot f_i(x_n) I(\delta = H_i)\, dx_1\ldots dx_n$$

$$= \sum_{i=1}^{k} \xi(i) \underbrace{\int f_i(x_1)\cdot\ldots\cdot f_i(x_n)\, dx_1\ldots dx_n}_{\text{integral of joint density}=1 \text{ and } \sum_{i=1}^{k}\xi(i)=1} - \int \sum_{i=1}^{k} \xi(i) f_i(x_1)\cdot\ldots\cdot f_i(x_n)\cdot I(\delta = H_i)\, dx_1\ldots dx_n$$

$$= 1 - \int \sum_{i=1}^{k} \xi(i) f_i(x_1)\cdot\ldots\cdot f_i(x_n)\cdot I(\delta = H_i)\, dx_1\ldots dx_n$$

So, to minimize Bayes error, we need to maximize $\int \sum_{i=1}^{k} \xi(i) f_i(x_1)\cdot\ldots\cdot f_i(x_n) I(\delta = H_i)\, dx_1\ldots dx_n$

i.e. we need to maximize the sum

$$\xi(1) f_1(x_1)\cdot\ldots\cdot f_1(x_n) I(\delta = H_1) + \ldots + \xi(k) f_k(x_1)\cdot\ldots\cdot f_k(x_n) I(\delta = H_k). \quad (*)$$

by choosing δ appropriately.

For each $(x_1,\ldots,x_n)$, δ chooses exactly one H_j, so all but one term in $(*)$ are zero. So, the contribution will be $\xi(j) f_j(x_1)\cdot\ldots\cdot f_j(x_n)$ for some j. Hence, in order to minimize Bayes error, δ should choose j that maximize this! $\square$

<u>Example</u> $k=2$ $\quad H_1: \mathbb{P}=\mathbb{P}_1$

$\quad\quad\quad\quad\quad\quad H_2: \mathbb{P}=\mathbb{P}_2$

Denote: $f_1(x_1,\dots,x_n)=f_1(x_1)\cdot\ldots\cdot f_1(x_n)$ joint p.d.f. of $(x_1,\dots,x_n)$ if $\mathbb{P}=\mathbb{P}_1$

$f_2(x_1,\dots,x_n)=f_2(x_1)\cdot\ldots\cdot f_2(x_n)$ ——— || ——— if $\mathbb{P}=\mathbb{P}_2$

Bayes error:

$$\alpha=\xi(1)\mathbb{P}_1(\delta\neq H_1)+\xi(2)\mathbb{P}_2(\delta\neq H_2)$$

Our test is:

$$\delta=\begin{cases} H_1, \text{ if } \xi(1)f_1(x_1,\dots,x_n)>\xi(2)f_2(x_1,\dots,x_n)\\ H_2, \text{ if } \xi(2)f_2(x_1,\dots,x_n)>\xi(1)f_1(x_1,\dots,x_n)\\ H_1 \text{ or } H_2, \text{ if } \xi(1)f_1(x_1,\dots,x_n)=\xi(2)f_2(x_1,\dots,x_n)\end{cases}$$

i.e.

$$\delta=\begin{cases} H_1, \text{ if } \dfrac{f_1(x_1,\dots,x_n)}{f_2(x_1,\dots,x_n)}>\dfrac{\xi(2)}{\xi(1)}\\ H_2, \text{ if } \dfrac{f_1(x_1,\dots,x_n)}{f_2(x_1,\dots,x_n)}<\dfrac{\xi(2)}{\xi(1)}\\ H_1 \text{ or } H_2, \text{ if } \dfrac{f_1(x_1,\dots,x_n)}{f_2(x_1,\dots,x_n)}=\dfrac{\xi(2)}{\xi(1)}\end{cases}$$

This type of test is usually called the <u>likelihood ratio test</u> since it's based on the quotient $\dfrac{f_1(x_1,\dots,x_n)}{f_2(x_1,\dots,x_n)}$ of likelihood functions.

So, suppose now that we have a sample $(X_1,\dots,X_n)$ and two hypotheses:

$H_1: \mathbb{P}=N(0,1)$

$H_2: \mathbb{P}=N(1,1)$ and arbitrary a priori weights $\xi(1)$ and $\xi(2)$

The likelihood ratio is:
$$\frac{f_1(x_1,\dots,x_n)}{f_2(x_1,\dots,x_n)}=\frac{\frac{1}{(\sqrt{2\pi})^n}\cdot e^{-\frac{1}{2}\sum_{i=1}^n x_i^2}}{\frac{1}{(\sqrt{2\pi})^n}\cdot e^{-\frac{1}{2}\sum_{i=1}^n (x_i-1)^2}}$$

which can be simplified to $e^{\frac{1}{2}\sum_{i=1}^{n}((X_i-1)^2-X_i^2)} = e^{\frac{n}{2}-\sum_{i=1}^{n}X_i}$

So, our test is:

$$\delta = \begin{cases} H_1, \text{ if } \sum_{i=1}^{n} X_i < \frac{n}{2} - \log\frac{\xi(2)}{\xi(1)} \\ H_2, \text{ if } \sum_{i=1}^{n} X_i > \frac{n}{2} - \log\frac{\xi(2)}{\xi(1)} \\ H_1 \text{ or } H_2, \text{ if } \sum_{i=1}^{n} X_i = \frac{n}{2} - \log\frac{\xi(2)}{\xi(1)} \end{cases}$$

6 Most powerful tests (Neyman-Pearson) for $k=2$

Sometimes, an error of type 1 is more important than the error of type 2. So, often, instead of minimizing the weighted (Bayes) error, we want to construct a test with CONTROLLED error of type 1 that minimizes error of type 2. So, we fix $\alpha \in (0,1)$ and consider only tests in the following class

$$K_\alpha = \{\delta : \mathbb{P}_1(\delta \neq H_1) \leq \alpha\}$$ and we try to find $\delta \in K_\alpha$

that minimizes $\alpha_2 = \mathbb{P}_2(\delta \neq H_2)$.

In some cases, this is easy!

Theorem If there exists c s.t. $\mathbb{P}_1\left(\frac{f_1(x_1,\ldots,x_n)}{f_2(x_1,\ldots,x_n)} < c\right) = \alpha$, then the test

$$\delta = \begin{cases} H_1, \text{ if } \frac{f_1(x_1,\ldots,x_n)}{f_2(x_1,\ldots,x_n)} \geq c \\ H_2, \text{ otherwise} \end{cases}$$

is the most powerful test in class K_α.

proof: The idea is to reduce this to Theorem 1 from Section 5. So, set $\xi(1) = \frac{1}{1+c}$ and $\xi(2) = \frac{c}{1+c}$. Then $\xi(1)+\xi(2) = 1$ and $\frac{\xi(2)}{\xi(1)} = c$. So, the test δ becomes $\delta = \begin{cases} H_1, \text{ if } \frac{f_1(x_1,\ldots,x_n)}{f_2(x_1,\ldots,x_n)} \geq \frac{\xi(2)}{\xi(1)} \\ H_2, \text{ otherwise} \end{cases}$

which is equivalent to the Bayes decision test in Theorem 1 of Section 5.

(see example in Section)

Note: The only difference is that here we break the tie in favor of H_1 always.

So, our test δ minimizes the Bayes error for these a priori weights $\xi(1)$ and $\xi(2)$.

In other words, for any other test δ', we have

$$\xi(1)\mathbb{P}_1(\delta\neq H_1)+\xi_2\mathbb{P}_2(\delta\neq H_2)\leq\xi(1)\mathbb{P}_1(\delta'\neq H_1)+\xi(2)\mathbb{P}_2(\delta'\neq H_2) \quad (*)$$

Is $\delta\in K_\alpha$? Yes, since $\mathbb{P}_1(\delta\neq H_1)=\mathbb{P}_1\left(\frac{f_1(x_1,\dots,x_n)}{f_2(x_1,\dots,x_n)}<c\right)=\alpha$.

If $\delta'\in K_\alpha$, then $\mathbb{P}_1(\delta'\neq H_1)\leq\alpha$ and thus, from $(*)$, we have that

$\mathbb{P}_2(\delta\neq H_2)\leq\mathbb{P}_2(\delta'\neq H_2)$, which exactly means that δ is more powerful than any other test in K_α.

Example Given: a sample $X=(X_1,\dots,X_n)$ & two hypotheses $H_1: \mathbb{P}=N(0,1)$

$H_2: \mathbb{P}=N(1,1)$.

Let's find the most powerful test δ with the error of type 1

$$\alpha_1\leq\alpha=0.05$$

So, according to Theorem 1, we look for c s.t. $\mathbb{P}_1\left(\frac{f_1(x_1,\dots,x_n)}{f_2(x_1,\dots,x_n)}<c\right)=\alpha=0.05$

This becomes

$$\mathbb{P}_1\left(\sum_{i=1}^n x_i>\frac{n}{2}-\log c\right)=\alpha=0.05, \text{ i.e.}$$

$$\mathbb{P}_1\left(\frac{1}{\sqrt{n}}\sum_{i=1}^n x_i>c'=\frac{1}{\sqrt{n}}\left(\frac{n}{2}-\log c\right)\right)=\alpha=0.05$$

But, if H_1 holds, then $\mathbb{P}_1=N(0,1)$ and, hence, $Y=\frac{1}{\sqrt{n}}\sum_{i=1}^n X_i\sim N(0,1)$

$$\text{so,}\quad \mathbb{P}(Y>c')=0.05\Rightarrow c'=1.64$$

So, the most powerful test w/ level of significance $\alpha=0.05$ is

$$\delta=\begin{cases}H_1, \text{ if } \frac{1}{\sqrt{n}}\sum_{i=1}^n x_i\leq 1.64\\ H_2, \text{ if } \frac{1}{\sqrt{n}}\sum_{i=1}^n x_i>1.64\end{cases}$$

What is error of type 2?

$$\alpha_2 = \mathbb{P}_2(\delta \neq H_2) = \mathbb{P}_2\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} X_i \leq 1.64\right) =$$

$$= \mathbb{P}_2\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(X_i - 1) \leq 1.64 - \sqrt{n}\right)$$

we subtracted 1 from each X_i since under H_2, $X_i \sim N(1,1)$. Now $Y = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(X_i - 1) \sim N(0,1)$, under H_2,

so $\alpha_2 = \mathbb{P}(Y \leq 1.64 - \sqrt{n}) = N(0,1)(-\infty, 1.64 - \sqrt{n})$

For $n = 10$, we would get $\alpha_2 = \mathbb{P}(Y \leq 1.64 - \sqrt{10}) = 0.087$.

<u>Note:</u> It is not always true that there exists c s.t. $\mathbb{P}_1\left(\frac{f_1(x_1,\dots,x_n)}{f_2(x_1,\dots,x_n)} < c\right) = \alpha$.
What do we do then? This is especially the case with discrete distributions $\mathbb{P}_i$.
In this case, there is a way to randomly break ties so that the condition still holds. We will not go into this, as it has little practical significance.

$\boxed{7}$ Chi-squared goodness-of-fit test (chi2gof in Matlab)

<u>Setup</u>: Given an i.i.d. sample $(X_1, \dots, X_n)$ from <u>unknown</u> distribution $\mathbb{P}$.
We want to test hypotheses of the type: $H_0 : \mathbb{P} = N(1,2)$
$H_0 : \mathbb{P} \neq N(1,2)$

It is different from t-tests where we would assume that data comes from some normal distribution w/ unknown parameters.

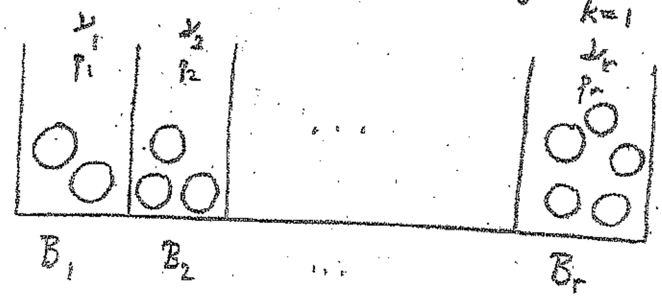
χ^2-goodness-of-fit test are based on <u>Pearson's theorem</u>.

<u>Setup</u>: - r boxes $B_1, \dots, B_r$

- throw n balls $X_1, \dots, X_n$ into these boxes independently of each other with probabilities $\mathbb{P}(X_i \in B_1) = p_1, \dots, \mathbb{P}(X_i \in B_r) = p_r$.

Of course, $\sum_{i=1}^{r} p_i = 1$.

Let ν_j be the ~~#of~~ # of balls in B_j, i.e. $\nu_j = \sum_{k=1}^{n} I(X_k \in B_j)$



$$\mathbb{E}(\nu_j) = \mathbb{E}\Big(\sum_{k=1}^{n} I(X_k \in B_j)\Big) \underset{\text{linearity of } \mathbb{E}}{=} \sum_{k=1}^{n} \mathbb{E}(I(X_k \in B_j)) = \sum_{k=1}^{n} \mathbb{P}(X_k \in B_j) = np_j.$$

So, by CLT, ν_j should be close to np_j.
However, we want something much stronger now: we want to describe the closeness of ν_j to np_j simultaneously for all boxes B_j, $j = \overline{1,r}$

Q: What's the difficulty? Random var's ν_j, $j = \overline{1,r}$ are not independent.
E.g. $\nu_1 + \dots + \nu_r = n$.

Theorem (Pearson) The random variable $\sum_{j=1}^{r} \frac{(\nu_j - np_j)^2}{np_j}$

converges in distribution to χ^2_{r-1}-distribution (as $n \to \infty$).

proof: Consider the indicator random var's

$$\mathbb{I}(X_1 \in B_j), \ldots, \mathbb{I}(X_n \in B_j) \quad \text{for the box } B_j.$$

They are i.i.d. Bernoulli $B(p_j)$ with probability of success

$$\mathbb{P}(X_i \in B_j) = \mathbb{E}(\mathbb{I}(X_i \in B_j)) = p_j \text{ and variance}$$

$$Var(\mathbb{I}(X_i \in B_j)) = p_j(1-p_j)$$

By $\overset{(\text{FACT } 3)}{\text{CLT}}$, we have:

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n}\mathbb{I}(X_i \in B_j) - \mathbb{E}\left(\mathbb{I}(X_i \in B_j)\right)\right) \xrightarrow{d} N(0, Var(\mathbb{I}(X_i \in B_j)))$$

i.e. $$\frac{\sum_{i=1}^{n}\mathbb{I}(X_i \in B_j) - np_j}{\sqrt{n}} \xrightarrow{d} N(0, p_j(1-p_j))$$

i.e. $$\frac{\nu_j - np_j}{\sqrt{np_j}} \xrightarrow{d} N(0, 1-p_j)$$

So, $\frac{\nu_j - np_j}{\sqrt{np_j}} \xrightarrow{d} Z_j$ where $Z_j \sim N(0, 1-p_j)$

Idea: We would like to know the distribution of $\sum_{j=1}^{r} Z_j^2$.

Correlation plays an important role, since ν_j's are not independent.

So, let's compute $Cov\left(\frac{\nu_i - np_i}{\sqrt{np_i}}, \frac{\nu_j - np_j}{\sqrt{np_j}}\right) = \mathbb{E}\left(\frac{\nu_i - np_i}{\sqrt{np_i}} \cdot \frac{\nu_j - np_j}{\sqrt{np_j}}\right) -$

$$- \underbrace{\mathbb{E}\left(\frac{\nu_i - np_i}{\sqrt{np_i}}\right)}_{0}\underbrace{\mathbb{E}\left(\frac{\nu_j - np_j}{\sqrt{np_j}}\right)}_{0} = \mathbb{E}\left(\frac{\nu_i - np_i}{\sqrt{np_i}} \cdot \frac{\nu_j - np_j}{\sqrt{np_j}}\right) =$$

$$= \frac{1}{n\sqrt{p_i p_j}}\left(E(\nu_i \nu_j) - np_j E(\nu_i) - np_i E(\nu_j) + n^2 p_i p_j\right) =$$

$$= \frac{1}{n\sqrt{p_i p_j}}\left(E(\nu_i \nu_j) - np_i np_j - np_j np_i + n^2 p_i p_j\right) = \frac{1}{n\sqrt{p_i p_j}}\left(E(\nu_i \nu_j) - n^2 p_i p_j\right)$$

So, we need to find $E(\nu_i \nu_j)$:

$$E(\nu_i \nu_j) = E\left(\left(\sum_{s=1}^{n} I(X_s \in B_i)\right)\left(\sum_{t=1}^{n} I(X_t \in B_j)\right)\right) = E\left(\sum_{s,t=1}^{n} I(X_s \in B_i) I(X_t \in B_j)\right)$$

$$\overset{\text{linearity of } E}{=} E\left(\sum_{s \neq t} I(X_s \in B_i) I(X_t \in B_j)\right) + E\left(\underbrace{\sum_{s=t} I(X_s \in B_i) I(X_t \in B_j)}_{\text{this is 0, since the same ball cannot be in two boxes at once}}\right)$$

$$= E\left(\sum_{s \neq t} I(X_s \in B_i) I(X_t \in B_j)\right) =$$

$$= n(n-1) E\left(I(X_s \in B_i) I(X_t \in B_j)\right) \overset{\text{independence}}{=} n(n-1) E\left(I(X_s \in B_i)\right) E\left(I(X_t \in B_j)\right)$$

$$= n(n-1) p_i p_j$$

So, $\mathrm{Cov}\left(\frac{\nu_i - np_i}{\sqrt{np_i}}, \frac{\nu_j - np_j}{\sqrt{np_j}}\right) = \frac{1}{n\sqrt{p_i p_j}}\left(n(n-1)p_i p_j - n^2 p_i p_j\right) = -\sqrt{p_i p_j}$

<u>Summary so far</u>

$$\sum_{j=1}^{r} \frac{(\nu_j - np_j)^2}{np_j} \xrightarrow{d} \sum_{j=1}^{r} Z_j^2, \text{ where } Z_j \sim N(0, 1-p_j)$$

and $E(Z_i Z_j) = -\sqrt{p_i p_j}$.

Now, we need to show that $\sum_{j=1}^{r} Z_j^2 \sim \chi^2_{r-1}$.

Let $\vec{g} = \begin{pmatrix} g_1 \\ \vdots \\ g_r \end{pmatrix}$, where $g_1 \dots g_r$ are i.i.d. $N(0,1)$, i.e. $\vec{g}$ is r-variate standard normal rand

Let $\vec{p} = \begin{pmatrix} \sqrt{p_1} \\ \vdots \\ \sqrt{p_r} \end{pmatrix}$. Consider a vector $\vec{g} - \langle \vec{g}, \vec{p} \rangle \vec{p}$ where

$\langle \vec{g}, \vec{p} \rangle = g_1\sqrt{p_1} + \dots + g_r\sqrt{p_r}$ is the usual scalar product of $\vec{g}$ and $\vec{p}$, i.e. $\vec{g}'\vec{p}$

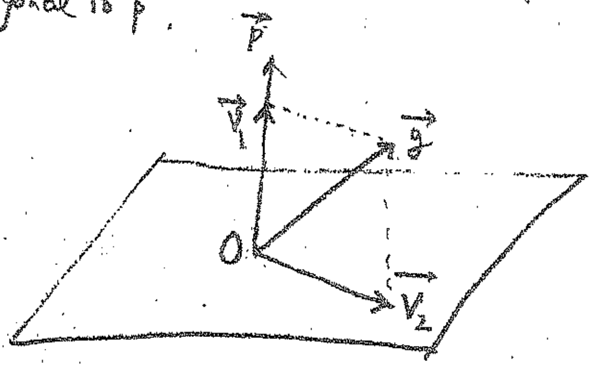
Claim $\vec{g} - \langle \vec{g}, \vec{p} \rangle \vec{p}$ has the same joint distribution as $(Z_1, \dots, Z_r)$

proof: Consider two coordinates, the i^{th}: $g_i - \left(\sum_{\ell=1}^{r} g_\ell \sqrt{p_\ell}\right)\sqrt{p_i}$ and the j^{th}: $g_j - \left(\sum_{\ell=1}^{r} g_\ell \sqrt{p_\ell}\right)\sqrt{p_j}$

Their covariance is $E\left[\left(g_i - \left(\sum g_\ell \sqrt{p_\ell}\right)\sqrt{p_i}\right)\left(g_j - \left(\sum g_\ell \sqrt{p_\ell}\right)\sqrt{p_j}\right)\right] = E[g_i g_j] - E\left[\left(\sum g_\ell \sqrt{p_\ell}\right) g_i \sqrt{p_j}\right] - E\left[\left(\sum g_\ell \sqrt{p_\ell}\right) g_j \sqrt{p_i}\right] + E\left[\sqrt{p_i}\sqrt{p_j}\left(\sum g_\ell \sqrt{p_\ell}\right)^2\right] = -\sqrt{p_i}\sqrt{p_j} - \sqrt{p_j}\sqrt{p_i} + \sqrt{p_i}\sqrt{p_j}\left(\sum_{\ell=1}^{r} p_\ell\right) = -\sqrt{p_i p_j}$

So,
$$\sum_{j=1}^{r} \left(\frac{\nu_j - np_j}{\sqrt{np_j}}\right)^2 \xrightarrow{d} \|\vec{g} - \langle \vec{g}, \vec{p} \rangle \vec{p}\|^2$$

But, what is $\vec{g} - \langle \vec{g}, \vec{p} \rangle \vec{p}$ geometrically? Well, $\|\vec{p}\| = \sum_{i=1}^{r} (\sqrt{p_i})^2 = \sum_{i=1}^{r} p_i = 1$ so $\vec{p}$ is a unit vector, $\vec{V_1} = \langle \vec{g}, \vec{p} \rangle \vec{p}$ is the projection of $\vec{g}$ onto the line along $\vec{p}$, and finally $\vec{V_2} = \vec{g} - \langle \vec{g}, \vec{p} \rangle \vec{p}$ is the projection of $\vec{g}$ onto the hyperplane orthogonal to $\vec{p}$.



Similarly, it is easy to compute that
$$E\left[\left(g_i - \left(\sum_{\ell=1}^{r} g_\ell \sqrt{p_\ell}\right)\sqrt{p_i}\right)^2\right] = 1 - p_i$$
□

Let's consider a new orthonormal coordinate system with the 1^{st} basis vector (i.e. axis) equal to $\vec{p}$. In this new coordinate system vector $\vec{g}$ has new coordinates $\tilde{\vec{g}} = (\tilde{g}_1, \tilde{g}_2, \dots, \tilde{g}_r) = V\vec{g}$ obtained from $\vec{g}$ by orthogonal transform
$$V = \begin{pmatrix} \vec{p}\,' \\ \vdots \end{pmatrix}$$
(i.e. first row of V is $\vec{p}\,'$) that maps the old basis into the new one.

Now, $\tilde{g}_1, \dots, \tilde{g}_r$ are also i.i.d. $N(0,1)$.

In the new coordinate system vector $\vec{V_2} = \vec{g} - \langle \vec{g}, \vec{p} \rangle \vec{p}$ has new coordinates $(0, \tilde{g}_2, \dots, \tilde{g}_r)'$, so
$$\|\vec{V_2}\|^2 = \|\vec{g} - \langle \vec{g}, \vec{p} \rangle \vec{p}\|^2 = (\tilde{g}_2)^2 + \dots + (\tilde{g}_r)^2$$
i.e. $\|\vec{V_2}\|^2$ has χ^2_{r-1}-distribution.

□

OK! So, suppose we have an i.i.d. sample $(X_1, \dots, X_n)$ of random variables that take a finite # of values $B_1, \dots, B_r$ with unknown probabilities $p_1, \dots, p_r$.

Consider a hypothesis:
$$\begin{cases} H_1: & p_i = p_i^0, \text{ for all } i=\overline{1,r} \\ H_2: & p_i \neq p_i^0 \text{ for some } i \end{cases}$$

If H_1 is true, then by Pearson's theorem
$$T = \sum_{i=1}^{r} \frac{(\nu_i - np_i^0)^2}{np_i^0} \xrightarrow{d} \chi^2_{r-1}.$$

Where $\nu_i = \#\{X_j : X_j = B_i\}$ are the observed counts in each category.

On the other hand, if H_2 holds, then for some i, $p_i \neq p_i^0$ and statistics T will behave differently. How differently?

Well, by CLT: $\dfrac{\nu_i - np_i}{\sqrt{np_i}} \xrightarrow{d} N(0, 1-p_i)$ and thus

$$\frac{\nu_i - np_i^0}{\sqrt{np_i^0}} = \frac{\nu_i - np_i + n(p_i - p_i^0)}{\sqrt{np_i^0}} = \sqrt{\frac{p_i}{p_i^0}}\,\frac{\nu_i - np_i}{\sqrt{np_i}} + \sqrt{n}\,\frac{p_i - p_i^0}{\sqrt{p_i^0}}$$

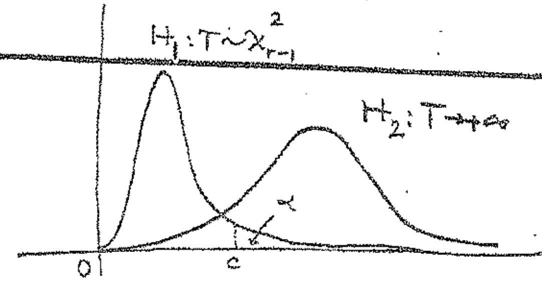
this goes to $\pm\infty$

this converges to $N\left(0, (1-p_i)\cdot\dfrac{p_i}{p_i^0}\right)$

So, if H_2 holds, then $\dfrac{(\nu_i - np_i^0)^2}{np_i^0} \to +\infty$ for some i, and therefore $T \to +\infty$.

<u>Conclusion</u>: As $n \to \infty$, the distribution of T under null hypothesis H_1 will approach χ^2_{r-1}-distr; and under H_2 it will shift to $+\infty$.

So, we define our test as follows:



$$\delta = \begin{cases} H_1 : T \leq c \\ H_2 : T > c \end{cases}$$

c is chosen so that

$$\alpha = \mathbb{P}_1(\delta \neq H_1) = \mathbb{P}_1(T > c) \approx \chi^2_{r-1}(c, \infty) \quad ; \text{ i.e. } \alpha = \chi^2_{r-1}(c, \infty)$$

δ is called the <u>χ^2-goodness-of-fit test</u>.

<u>Example</u>: In 89, 189 Zimbabwe residents were polled whether their financial status was better, worse or same than 1 year ago.

worse	same	better	total
58	64	67	189

We want to test the hypothesis that the underlying distribution is uniform, i.e. $H_1 : p_1 = p_2 = p_3 = 1/3$. Take $\alpha = 0.05$ as the level of significance.
c is found from $\chi^2_{3-1}(c, \infty) = 0.05$, which gives $c = 5.9$
The chi-squared statistic is

$$T = \frac{(58 - 189/3)^2}{189/3} + \frac{(64 - 189/3)^2}{189/3} + \frac{(67 - 189/3)^2}{189/3} = 0.666 < 5.9$$

so, we accept H_1 at the level of significance 0.05.

<u>Question</u>: All of this works for distributions with finite # of values in their range. What about continuous distributions? We perform <u>discretization</u>.
For example, consider a random variable X which takes values in the interval $0 < X < 1$ but has an unknown p.d.f. on this interval.
Suppose that a random sample of 100 observations is taken from this unknown distribution, and we want to test the null hypotheses that the distribution is uniform on $(0,1)$. What do we do?

∴ Divide the interval $(0,1)$ into 20 subintervals of equal length: $(0, 0.05)$, $(0.05, 0.1)$ etc.

∴ If the actual distribution is uniform, then the probability that any particular observation falls into the i^{th} interval is $1/20$, $i=\overline{1,20}$.

Since the sample size is $n=100$, then the expected # of observations in each subinterval is 5. If ν_i is the # of observations in the sample that actually fall within the i^{th} subinterval, then the statistic T is

$$T = \sum_{i=1}^{20} \frac{(\nu_i - 5)^2}{5}$$

If the null hypothesis is true, then T will have approximately a χ^2_{19}-distribution.

This method can be applied to any continuous distribution.

<u>Step 1</u>: partition the real line into a finite # of intervals, r.

Sometimes, this is done so that the expected count in each interval $np_i^0 \geq 5$ (Matlab does this). Here p_i^0 is the probability that the particular hypothesized distribution would assign to the i^{th} interval.

(Sometimes the intervals are defined so that $p_i^0 = \frac{1}{r}$ for all i and r is chosen so that $np_i^0 = \frac{n}{r} \geq 5$)

<u>Step 2</u> Count ν_i: # of observations in the sample that fall within the i^{th} subinterval

<u>Step 3</u>. Calculate $T = \sum_{i=1}^{r} \frac{(\nu_i - np_i^0)^2}{np_i^0}$

If the null hypothesis is true, then $T \approx \chi^2_{r-1}$.

8 χ^2-goodness of fit tests for composite hypotheses

- Setup: Before – given a sample $(X_1, \dots, X_n)$ from unknown distribution, we want to test whether the distribution is some specific distribution, e.g. $U(0,1)$.
Now, we want to test whether data comes from a FAMILY of distributions (this is called a composite hypothesis). We will use χ^2-g.o.f. tests again!

8.1 Discrete distributions

In this case, a random variable (i.e. data point, i.e. observation) takes a finite number of values $B_1, \dots, B_r$ with probabilities

$$p_1 = \mathbb{P}(X=B_1),\ p_2 = \mathbb{P}(X=B_2), \dots, p_r = \mathbb{P}(X=B_r)$$

Based on a given sample $(X_1, \dots, X_n)$, we would like to test a hypothesis that this sample comes from a family of distributions $\{\mathbb{P}_\theta : \theta \in \Omega\}$.
Denote $p_j(\theta) = \mathbb{P}_\theta(X=B_j)$, $j=\overline{1,r}$. Then, we want to test

$$\begin{cases} H_0 : p_j = p_j(\theta) & \text{for all } j=\overline{1,r}, \text{ for some } \theta \in \Omega \\ H_1 : \text{otherwise} \end{cases}$$

What we did in Section 7 is the case when Ω has a single possible value for θ.
We know that in this case we would use the simple χ^2-g.o.f. test based on the statistic

$$T = \sum_{j=1}^{r} \frac{(\nu_j - np_j(\theta))^2}{np_j(\theta)}$$

Now, the situation is more complicated, since there are many candidates for θ.
One way to approach this is as follows:

Step 1. Assuming that H_0 holds (i.e. $\mathbb{P}=\mathbb{P}_\theta$ for some $\theta \in \Omega$), we find an estimate θ^* of the unknown θ. It turns out that a good choice for θ^* is the MLE of θ, i.e. the value of θ that maximizes the likelihood function

$$\varphi(\theta) = p_1(\theta)^{\nu_1} \cdots p_r(\theta)^{\nu_r} \quad \text{(or its log)}$$

Step 2. Try to test if, indeed, the distribution $\mathbb{P}$ is equal to $\mathbb{P}_{\theta^*}$ by using the statistic

$$T=\sum_{j=1}^{r}\frac{(\nu_j-np_j(\theta^*))^2}{np_j(\theta^*)}$$

and the usual χ^2-g.o.f. test.

It turns out that if θ^* is the MLE for θ, then

$$T=\sum_{j=1}^{r}\frac{(\nu_j-np_j(\theta^*))^2}{np_j(\theta^*)}\xrightarrow{d}\chi^2_{r-s-1}$$

where s is the "dimension" of the parameter set Ω.

(We'll see in an example below how to determine dimension of Ω)

Step 3. Our test is

$$\delta=\begin{cases}H_0: & T\le c\\ H_1: & T>c\end{cases}$$

where c is determined from

$$\alpha=\mathbb{P}(\delta\ne H_0\,|\,H_0)=\mathbb{P}(T>c\,|\,H_0)\approx\chi^2_{r-s-1}(c,\infty).$$

Example (binomial distribution $B(2,\theta)$) A gene has 2 alleles A_1 and A_2 and the combinations of these alleles define three genotypes A_1A_1, A_1A_2 and A_2A_2. We want to test a theory that

$$H_0:\begin{cases}P_1(\theta)=\mathbb{P}(A_1A_1)=\theta^2\\ P_2(\theta)=\mathbb{P}(A_1A_2)=2\theta(1-\theta)\\ P_3(\theta)=\mathbb{P}(A_2A_2)=(1-\theta)^2\end{cases}$$

In other words, we want to test the theory that the probability to pass on A_1 is θ and the — " — A_2 is $1-\theta$ for some $0<\theta<1$.

Suppose we're given a random sample $X_1,\dots,X_{353}$ from the population with counts of each genotype: $\nu_1=35$, $\nu_2=160$, $\nu_3=158$

Step 1. Find the MLE θ^* of θ, assuming H_0 is true.

So, we are maximizing $\varphi(\theta)=p_1(\theta)^{\nu_1}p_2(\theta)^{\nu_2}p_3(\theta)^{\nu_3}$, or equivalently,

$$\therefore \log p_1(\theta)^{\nu_1} p_2(\theta)^{\nu_2} p_3(\theta)^{\nu_3} = \nu_1 \log p_1(\theta) + \nu_2 \log p_2(\theta) + \nu_3 \log p_3(\theta) =$$
$$= \nu_1 \log \theta^2 + \nu_2 \log 2\theta(1-\theta) + \nu_3 \log (1-\theta)^2.$$

Find the derivative $\frac{d}{d\theta}$ & set to $0 \Rightarrow$

$$\frac{2\nu_1}{\theta} + \nu_2 \cdot \frac{2-4\theta}{2\theta(1-\theta)} - \frac{2\nu_3}{1-\theta} = 0 \Rightarrow 4\nu_1(1-\theta) + \nu_2(2-4\theta) - 4\nu_3\theta = 0$$
$$\Rightarrow \theta(4\nu_1 + 4\nu_2 + 4\nu_3) = 4\nu_1 + 2\nu_2 \Rightarrow 4\theta \cdot n = 4\nu_1 + 2\nu_2 \Rightarrow$$
$$\Rightarrow \theta^* = \frac{2\nu_1 + \nu_2}{2n} = 0.325779$$

Step 2. Under H_0, $T = \frac{(\nu_1 - np_1(\theta^*))^2}{np_1(\theta^*)} + \frac{(\nu_2 - np_2(\theta^*))^2}{np_2(\theta^*)} + \frac{(\nu_3 - np_3(\theta^*))^2}{np_3(\theta^*)}$

should converge in distribution to $\chi^2_{r-s-1} = \chi^2_{3-1-1} = \chi^2_1$

since $s = 1$ (distribution is determined by one parameter, ~~namely~~ θ)

Calculating ... $T = 0.052819 + 0.156683 + 0.037854 = 0.247356$.

Step 3. Our test is

$$\delta = \begin{cases} H_0: T \le c \\ H_1: T > c \end{cases}$$ where c comes from $\alpha = 0.05 = \chi^2_1(c, \infty)$, i.e. $c = 3.841$

Now $T \le c$, so we accept the hypothesis H_0.

8.2 Continuous distributions

What if the distributions $\mathbb{P}_\theta$, $\theta \in \Omega$ are continuous distributions?

Our hypothesis: $H_0: \mathbb{P} = \mathbb{P}_\theta$ for some $\theta \in \Omega$.

Group the data into r intervals $I_1, \dots, I_r$ and instead test the hypothesis:

$$H_0: p_j = p_j(\theta) = \mathbb{P}_\theta(X \in I_j) \text{ for all } j = \overline{1,r} \text{ for some } \theta \in \Omega$$

Example: Discretize normal distribution by grouping the data into r intervals $I_1, \dots, I_r$.
The hypothesis is:

$$H_0: p_j = N(\mu, \sigma^2)(I_j) \text{ for all } j = \overline{1,r} \text{ for some } (\mu, \sigma^2)$$

There are 2 parameters (μ and σ^2) that describe all probabilities, so $s=2$.
Our statistic T should converge to $\chi^2_{r-s-1} = \chi^2_{r-3}$-distribution.

Note: It can be difficult to maximize the "grouped" likelihood function

$$\mathbb{P}_\theta(I_1)^{\nu_1} \dots \mathbb{P}_\theta(I_r)^{\nu_r} \quad \text{in order to get } \theta^*$$

It's tempting to use a usual non-grouped MLE $\hat{\theta}$ of θ instead, because we know explicit formulae for the MLE's of many distributions.
However, the statistic

$$T = \sum_{j=1}^{r} \frac{(\nu_j - np_j(\hat{\theta}))^2}{np_j(\hat{\theta})} \quad \text{no longer}$$

converges to χ^2_{r-s-1}-distribution.

A famous result of Chernoff & Lehmann states that T DOES CONVERGE to some distribution "in between" χ^2_{r-1} and χ^2_{r-s-1}.

8.3. Asymptotic inference with the MLE

As in Section 2, let $X_1, \dots, X_n$ be an i.i.d. sample from a family of distributions $\{\mathbb{P}_\theta : \theta \in \Omega\}$
Let $\ell(\theta) = \sum_{i=1}^{n} \log f_\theta(X_i)$ be the log-likelihood of this sample and let $\hat{\theta}$ denote the MLE.
By FACT 11 (page 24.), we have: $(\hat{\theta} - \theta_0)'\left(-(\nabla^2 \ell(\theta))\big|_{\theta=\hat{\theta}}\right)(\hat{\theta} - \theta_0) \xrightarrow{d} \chi^2_p$ as $n \to \infty$.
One can use this asymptotic behaviour to set up a $100(1-\alpha)\%$ confidence ellipsoid for θ_0, the p-dimensional vector containing the true values of the unknown parameters θ.
The $100(1-\alpha)\%$ confidence ellipsoid is: $\left\{\theta : (\hat{\theta} - \theta)'\left(-(\nabla^2 \ell(\theta))\big|_{\theta=\hat{\theta}}\right)(\hat{\theta} - \theta) \leq \chi^2_{p;1-\alpha}\right\}$
where α is usually set to 0.05 and $\chi^2_{p;1-\alpha}$ denotes the $(1-\alpha)$-quantile of χ^2_p-distribution.
So, approximately speaking, θ_0 is inside this ellipsoid with probability $\geq 1-\alpha$, as $n \to \infty$.

One can also test the hypothesis of the form $H_0: \theta \in \Omega_0$, where Ω_0 is a q-dimensional subspace of the parameter space Ω $(0 \leq q < p)$ using the generalized likelihood ratio (GLR) statistic $\Lambda := 2\left\{\ell(\hat{\theta}) - \sup_{\theta \in \Omega_0} \ell(\theta)\right\}$. It is known that, under H_0, Λ has a limiting χ^2_{p-q} distribution (as $n \to \infty$). Hence, the GLR test with significance level α rejects H_0 if Λ exceeds $\chi^2_{p-q;1-\alpha}$.

9 Test of independence

Setup: Our observations are classified by two different features and we would like to test if these features are independent.

Each observation X is of the type (i,j): $i=\overline{1,a}$, $j=\overline{1,b}$.

(i,j): i — 1st feature, j — 2nd feature. notation: $X^1=i$, $X^2=j$.

Contingency table for an i.i.d. sample $(X_1,\dots,X_n)$:

	Feature 2			
Feature 1	1	2	...	b
1	N_{11}	N_{12}	...	N_{1b}
2	N_{21}	N_{22}	...	N_{2b}
⋮	⋮	⋮	⋮	⋮
a	N_{a1}	N_{a2}		N_{ab}

N_{ij} = # of observations in cell (i,j).

Let $\mathbb{P}(X=(i,j))=\theta_{ij}$, $\mathbb{P}(X^1=i)=p_i$, $\mathbb{P}(X^2=j)=q_j$.

Then, our hypotheses are:

$$\begin{cases} H_0: \theta_{ij}=p_iq_j \text{ for all } (i,j) \text{ for some } (p_1,\dots,p_a) \text{ and } (q_1,\dots,q_b) \\ H_1: \text{otherwise} \end{cases}$$

The null hypotheses H_0 is a special case of the composite hypotheses from Section 8. So, the idea is to use chi-squared goodness-of-fit test.

What is the dimension of the parameter set?

$p_1+\dots+p_a=1$ and $q_1+\dots+q_b=1$, so

we can take $(p_1,\dots,p_{a-1})$ and $(q_1,\dots,q_{b-1})$ as free parameters of the model.

So, the dimension of the parameter set is

$$s=(a-1)+(b-1)=a+b-2.$$

Therefore, if we find the MLE's for the parameters of this model, then

$$T = \sum_{i,j} \frac{(N_{ij} - np_i^* q_j^*)^2}{np_i^* q_j^*} \xrightarrow{d} \chi^2_{r-s-1} = \chi^2_{(a-1)(b-1)}$$

since $r = a \cdot b$ is the # of groups and $s = a + b - 2$.

To formulate the test, it remains to find the MLE's of the parameters.

We need to maximize the likelihood function

$$\prod_{i,j} (p_i q_j)^{N_{ij}} = \prod_i p_i^{\sum_j N_{ij}} \prod_j q_j^{\sum_i N_{ij}} = \prod_i p_i^{N_{i+}} \cdot \prod_j q_j^{N_{+j}}$$

where we introduced notation $N_{i+} = \sum_j N_{ij}$ and $N_{+j} = \sum_i N_{ij}$ for the total # of observations in the i^{th} row and j^{th} column, respectively.

p_i's and q_j's are not related, so maximizing the likelihood function above is equivalent to maximizing $\prod_i p_i^{N_{i+}}$ and $\prod_j q_j^{N_{+j}}$ separately.

So, let's maximize $\prod_i p_i^{N_{i+}}$, or taking the logarithm, maximize

$$\sum_{i=1}^{a} N_{i+} \log p_i = \sum_{i=1}^{a-1} N_{i+} \log p_i + N_{a+} \log(1 - p_1 - \ldots - p_{a-1})$$

$$\frac{d}{dp_i} \sum_{i=1}^{a} N_{i+} \log p_i = 0 \Rightarrow \frac{N_{i+}}{p_i} - \frac{N_{a+}}{1 - p_1 - \ldots - p_{a-1}} = \frac{N_{i+}}{p_i} - \frac{N_{a+}}{p_a} = 0$$

$\Rightarrow \boxed{N_{i+} p_a = N_{a+} p_i}$ Add these equations for all $i = \overline{1,a}$, we get

$$n p_a = N_{a+} \Rightarrow p_a = \frac{N_{a+}}{n} \Rightarrow p_i = \frac{N_{i+}}{n}$$

Therefore, the MLE for p_i, $i = \overline{1,a}$ is $p_i^* = \frac{N_{i+}}{n}$

Similarly, the MLE for q_j, $j = \overline{1,b}$ is $q_j^* = \frac{N_{+j}}{n}$

So, the chi-square statistic T is:

$$T = \sum_{i,j} \frac{\left(N_{ij} - \frac{N_{i+}N_{+j}}{n}\right)^2}{\frac{N_{i+}N_{+j}}{n}}$$ and our test is

$$\delta = \begin{cases} H_1 : T \le c \\ H_2 : T > c \end{cases}$$

where the threshold c comes from $\alpha = \chi^2_{(a-1)\cdot(b-1)}(c, +\infty)$, or in a quantile notation $c = \chi^2_{(a-1)(b-1);1-\alpha}$.

<u>Example</u> In 2008, Zimbabwe poll: ? personal financial status better than 1 yr ago?

Income \ Opinion	$b=3$			
$a=3$	Worse	Same	Better	
$\le 20K$	20	15	12	47
$\ge 20K, \le 35K$	24	27	32	83
$\ge 35K$	14	22	23	59
	58	64	67	189

$$T = \frac{(20 - 47\times 58/189)^2}{47\times 58/189} + \dots + \frac{(23 - 67\times 59/189)^2}{67\times 59/189} = 5.21.$$

If we take $\alpha = 0.05$, then

$$0.05 = \chi^2_{(a-1)(b-1)}(c, +\infty) = \chi^2_4(c, \infty) \Rightarrow c = 9.488$$

$T = 5.21 < c = 9.488$, so we accept the null hypothesis that opinions are independent of income.

10. Kolmogorov – Smirnov Test

(kstest in Matlab)
(lilliefest in Matlab)
(for Lilliefor's ≡ adjusted K-S test)

Given an i.i.d. sample $X_1, X_2, \dots, X_n$ with some UNKNOWN distribution $\mathbb{P}$, we would like to test the hypothesis that $\mathbb{P}$ is equal to a particular distribution $\mathbb{P}_0$, i.e. decide between the hypotheses:

$$\begin{cases} H_0: \mathbb{P} = \mathbb{P}_0 \\ H_1: \mathbb{P} \neq \mathbb{P}_0 \end{cases}$$

We already know how to test this hypothesis using χ^2-goodness-of-fit test. If distribution $\mathbb{P}_0$ is continuous, we had to group the data and consider a weaker discretized null hypothesis. We now consider a different test for H_0 based on a very different (perhaps even more natural) idea that AVOIDS DISCRETIZATION!

Let $F(x) = \mathbb{P}(X_1 \leq x)$ be a c.d.f. of a true underlying distribution of the data.

Define an EMPIRICAL C.D.F by:

$$F_n(x) = \mathbb{P}_n(X \leq x) = \frac{1}{n} \sum_{i=1}^{n} I(X_i \leq x)$$

Note: Some authors call this the sample c.d.f.

i.e. as the proportion of the sample points that are below level x.

Note that by the LLN (Law of large numbers), we have for every fixed $x \in \mathbb{R}$:

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} I(X_i \leq x) \longrightarrow E[I(X_1 \leq x)] = \mathbb{P}(X_1 \leq x) = F(x)$$

So $F_n(x) \longrightarrow F(x)$ as $n \to \infty$, i.e.

$$\forall x \in \mathbb{R},\ \forall \varepsilon > 0,\ \exists N(\varepsilon, x) \text{ s.t. for all } n \geq N: \ \mathbb{P}(|F_n(x) - F(x)| > \varepsilon) = 0$$

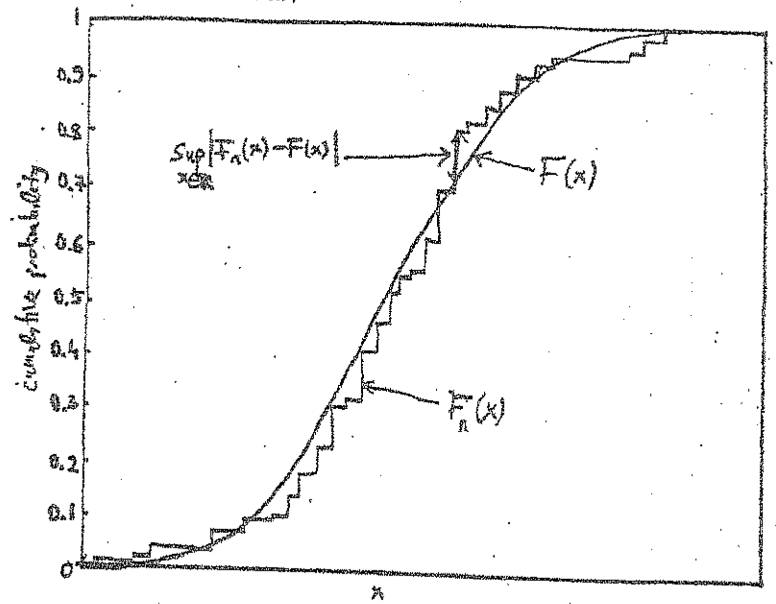
(so called pointwise convergence)

Also, by CLT, we have:

$\forall x \in \mathbb{R}: \sqrt{n}\,(F_n(x) - F(x)) \xrightarrow{d} N(0, F(x)(1-F(x)))$ (as $n \to \infty$), since $F(x)(1-F(x))$ is the variance of $I(X_1 \leq x)$.

Kolmogorov and Smirnov proved more (we will not go into this as the proof is out of ~~scope and difficult). They considered the following natural random variable:~~

$\sup_{x\in\mathbb{R}} |F_n(x)-F(x)|$



<u>Theorem</u> 1) If the unknown distribution $\mathbb{P}$ of the sample is continuous, then $\sup_{x\in\mathbb{R}} |F_n(x)-F(x)|$ does not depend on F (i.e. on $\mathbb{P}$).

2) $\forall \varepsilon>0$, $\exists N(\varepsilon)$ s.t. for all $n \geq N$: $\mathbb{P}\left(\sup_{x\in\mathbb{R}} |F_n(x)-F(x)| > \varepsilon\right) = 0$, i.e.

$\forall \varepsilon>0$, $\exists N(\varepsilon)$ s.t. for all $x\in\mathbb{R}$, and all $n \geq N$: $\mathbb{P}\left(|F_n(x)-F(x)| > \varepsilon\right) = 0$

(so called <u>uniform convergence</u>, which is much stronger than the pointwise convergence that follows from LLN)

3) The random variable $\sqrt{n}\sup_{x\in\mathbb{R}} |F_n(x)-F(x)|$ also converges in distribution as follows:

$$\lim_{n\to\infty} \mathbb{P}\left(\sqrt{n}\cdot\sup_{x\in\mathbb{R}} |F_n(x)-F(x)| \leq t\right) = H(t), \text{ where}$$

$$H(t) = 1 - 2\cdot\sum_{i=1}^{\infty}(-1)^{i-1} e^{-2i^2 t}$$ is the c.d.f. of Kolmogorov–Smirnov distribution.

Let's use the Theorem to set up the K-S test:

First, we reformulate our hypotheses in terms of the cumulative distribution functions:

$$\begin{cases} H_0: F = F_0 \\ H_1: F \neq F_0 \end{cases}, \text{ where } F_0 \text{ is the c.d.f. of } \mathbb{P}_0.$$

We consider the following statistic: $\boxed{D_n = \sqrt{n} \sup_{x\in\mathbb{R}} |F_n(x) - F_0(x)|}$

If the null hypothesis is true, then by Theorem, the distribution of D_n will depend only on n, and if n is large, the distribution of D_n is approximated by the Kolmogorov-Smirnov distribution.

If the null hypothesis is not true, i.e. if $F \neq F_0$, then since F is the true c.d.f. of the data, by LLN the empirical c.d.f. F_n converges to F as $n \to \infty$, and as a result will not approximate F_0, i.e. for large n we will have:

$$\sup_{x\in\mathbb{R}} |F_n(x) - F_0(x)| > \varepsilon \text{ for some small enough } \varepsilon.$$

This implies that $D_n = \sqrt{n} \sup_{x\in\mathbb{R}} |F_n(x) - F_0(x)| > \sqrt{n}\,\varepsilon$. So,

if H_0 fails, then $D_n > \sqrt{n}\,\varepsilon \to +\infty$ as $n \to \infty$.

Therefore, to test H_0 we will consider a decision rule:

$$\delta = \begin{cases} H_0, \text{ if } D_n \le c \\ H_1, \text{ if } D_n > c, \end{cases}$$

where the treshold c depends on the level of significance α:

$$\alpha = \mathbb{P}(\delta \neq H_0 | H_0) = \mathbb{P}(D_n > c | H_0)$$

Since under H_0, the distribution of D_n depends only on n and can be tabulated, we can find the $\underbrace{\text{treshold}}_{c=c_\alpha}$ from the tables. Most softwares/statistical tables have these distributions for $n \le 100$. Moreover, when n is large (as often in practice), we can use the KS distribution to find c, since

$$\alpha = \mathbb{P}(D_n \ge c / H_0) \approx 1 - H(c)$$

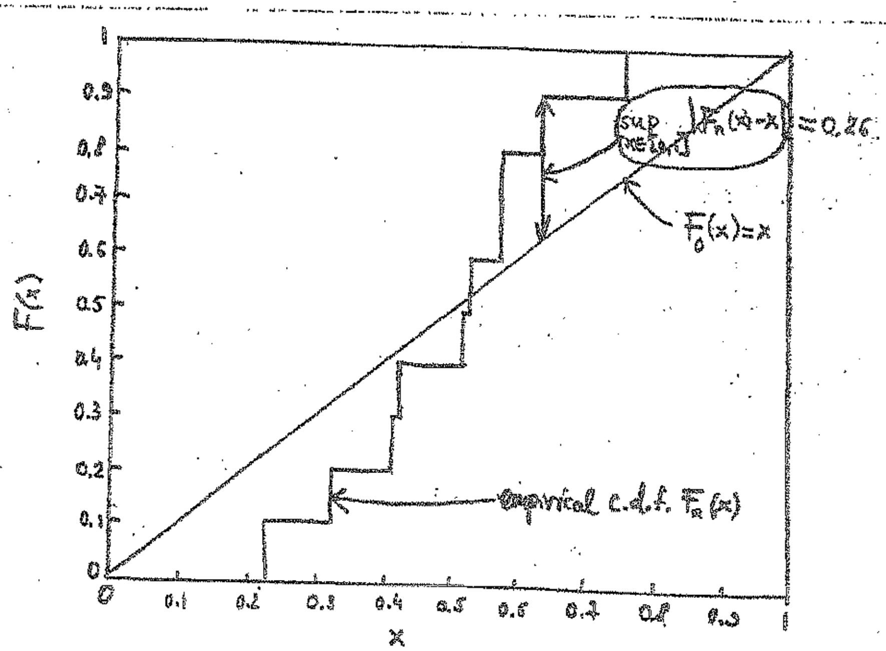
and we can use the tables for H to find c.

<u>Example</u> Consider a sample of size 10:

$$0.58, 0.42, 0.52, 0.33, 0.43, 0.23, 0.58, 0.76, 0.53, 0.64.$$

Let's test the hypothesis that the distribution of the sample is uniform on $[0,1]$, i.e.

$$H_0: \; F(x) = F_0(x) = x.$$



To compute D_n, notice that the largest difference between $F_0(x)$ and $F_n(x)$ is achieved either before or after one of the jumps, so calculate $|F_n(x) - F_0(x)|$ before and after each jump.

before the jump	after the jump
$\|0-0.23\|$	$\|0.1-0.23\|$
$\|0.1-0.23\|$	$\|0.2-0.33\|$
$\|0.2-0.42\|$	$\|0.3-0.42\|$
$\|0.3-0.43\|$	$\|0.4-0.43\|$
$\vdots$	$\vdots$

The largest value is $|0.9-0.64| = 0.26$, so $D_n = \sqrt{n} \sup_{x\in[0,1]} |F_n(x) - x| = \sqrt{10} \cdot 0.26 = 0.82$.

Now, close your eyes and pretend that $n=10$ is very large. Then the KS approximation from part 3) of Theorem gives: $1 - H(c) = \underset{\alpha}{\underset{\uparrow}{0.05}} \Rightarrow c = 1.35$

So, the KS test is:

$$\delta = \begin{cases} H_1, \text{if } D_n \le 1.35 \\ H_2, \text{if } D_n > 1.35 \end{cases}$$

So we accept the null hypothesis H_0, since $D_n = 0.82 < c = 1.35$.

<u>K-S test for two samples</u> (kstest2 in Matlab)

This is very similar. Suppose that a first sample $X_1, \dots, X_m$ of size m has distribution with c.d.f. $F(x)$ and the second sample $Y_1, Y_2, \dots, Y_n$ of size n has distribution with c.d.f. $G(x)$ and we want to test

$$\begin{cases} H_0 : F = G \\ H_1 : F \ne G. \end{cases}$$

Let $F_m(x)$ and $G_n(x)$ be the corresponding empirical c.d.f.'s. Then the statistic

$$D_{mn} = \left(\frac{mn}{m+n}\right)^{1/2} \sup_{x\in\mathbb{R}} |F_m(x) - G_n(x)|$$ satisfies the same

Theorem as before. The rest is the same.

10 Addendum: Practical measures of nonnormality

A widely popular and disputed hypothesis in financial markets is that the log returns on stocks are normally distributed. There are several ways to test this hypothesis in practice:

(A) Look at the normal probability plot of returns, also known as the Q–Q plots.

This is a plot of the sample quantiles versus the quantiles of the standard normal $N(0,1)$ distribution. We know what quantiles are; but, what are sample quantiles? Order the data $X_1, \dots, X_n$ from smallest to largest to get $X_{(1)}, X_{(2)}, \dots, X_{(n)}$; i.e. so called order statistics. The q-sample quantile (also known as the $100q$th sample percentile) is $X_{(k)}$, where $k = qn$ rounded to an integer. Some authors round up, some down, some round in both directions and use a weighted average of the two results.

If the normality assumption is true, then the q^{th} sample quantile will be approximately equal to $\mu + \sigma \Phi^{-1}(q)$, where Φ^{-1} denotes the inverse of the standard normal c.d.f.

In other words, aside for sampling variation, a plot of the sample quantiles versus the normal quantiles Φ^{-1} will be linear.

Another way to put it is that the normal probability plot is a plot of $X_{(i)}$ versus $\Phi^{-1}\left(\frac{i}{n+1}\right)$

Systematic deviation of the plot from a straight line is evidence of nonnormality!

(In Matlab normplot)

(B) Stock return distributions have often been observed to have heavy tails (which indicate the possibility of an extremely large negative return which could, for example, entirely deplete the capital reserves of a firm). Good measure for heavy tails is high kurtosis. Check the homework for definition of a kurtosis of a random variable.

Sample kurtosis is defined as $\hat{K} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{X_i - \bar{X}}{s}\right)^4$, where $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$ as before

Both the sample skewness $\left(\hat{Sk} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{X_i - \bar{X}}{s}\right)^3\right)$ and the excess sample kurtosis $(\hat{K} - 3)$ should be near zero if a sample is from a normal distribution. Caveat: be aware of possible outliers when using (B)

(C) Use normality tests such as Kolmogorov–Smirnov, Shapiro–Wilk, Anderson–Darling etc. Check the p-values. For example, procedure PROC UNIVARIATE of SAS does all this for you!