UNIVERZITA PALACKÉHO V OLOMOUCI
P Ř Í R O D O V Ě D E C K Á F A K U L T A
KATEDRA MATEMATICKÉ ANALÝZY A APLIKACÍ MATEMATIKY

BAKALÁŘSKÁ PRÁCE

Tutoriál statistických metod pro populační asociační studie

Vedoucí diplomové práce:
Mgr. Jana Vrbková
Rok odevzdání: 2010

Vypracovala:
Martina Vrzalová
ME, III. Ročník

Prohlášení

Prohlašuji, že jsem tuto bakalářskou práci vypracovala samostatně pod vedením Mgr. Jany Vrbkové, a že jsem v seznamu použité literatury uvedla všechny použité zdroje.

V Olomouci dne 29. 4. 2010

……

Poděkování

Ráda bych tímto poděkovala především své vedoucí diplomové práce paní Mgr. Janě Vrbkové, za její ochotu, trpělivost a čas, který mi věnovala v době konzultací a za psychickou podporu při časovém presu. Poděkování si zaslouží i má rodina, která mě po celý čas studia podporovala.

Obsah

Úvod

Tato práce pojednává o základních statistických procedurách užívaných v rámci populačních asociačních studií. Populační asociační studie zkoumají asociace mezi polymorfismy a výskytem nemoci a berou v úvahu do studie pouze jedince, mezi nimiž není prokázán příbuzenský stav.

Cílem této práce je poskytnout čtenáři přehled základních statistických metod užívaných v populačních asociačních studiích, se zaměřením na asociaci založenou na jednonukleotidovém polymorfismu (SNP). Vycházím zejména z knihy Applied Statistical Genetics with R For Population-based Association Studies, Andrey S. Foulkes. [1].

Kromě popisu jednotlivých statistických procedur se zaměřím i na samostatné zpracování dat s podporou statistických softwarů R a SAS. Jako příklad zpracování v systému R používám již publikované příklady v literatuře [1], které podrobněji komentuji, a tytéž příklady samostatně zpracovávám v softwaru SAS.

Pro pochopení mé práce předpokládám u čtenáře znalost statistických pojmů a alespoň základní znalost prostředí systémů R a SAS.

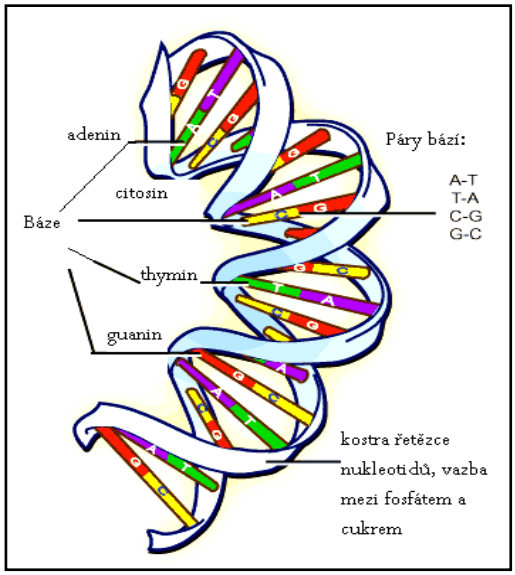
1 Populační asociační studie, základní pojmy

V současné době je na světě přes 6 miliard lidí, přesto nepotkáte více lidí, kteří vypadají stejně (kromě jednovaječných dvojčat či vícerčat). Nejen náš vzhled, ale i výskyt onemocnění a zdravotních potíží, je určován nespočetným množstvím faktorů, jak genetických, tak negenetických.

Genetika patří mezi biologické vědy a zabývá se dědičností a proměnlivostí živých soustav. Má mnoho specializačních oblastí a jednou z nich je právě populační asociační genetika, která je náplní této práce. Populační asociační studie se zaměřují na odhalení asociace mezi genotypem člověka a daným znakem (nemocí či jejím projevem). Jsou založeny na dvou stěžejních konceptech, konceptu Hardyho-Weinbergovy rovnováhy a vazebné nerovnováhy. Za použití vybraných statistických metod, za určitých předpokladů, asociační studie zkoumají, zda výskyt dané nemoci je podmíněn právě genetickou změnou v sekvenci DNA. Důležitou úlohou populačních asociačních studií je také pečlivě zvážit specifika pacienta (proměnné), která by mohla být potenciálními zavádějícími faktory nebo modifikátory. Zvážení vlivu těchto faktorů na vztah mezi genotypem a znakem přispívá ke správným závěrům studie.

1.1 DNA, geny, znaky

DNA (kyselina deoxyribonukleová) je zobrazována jako dvojitá spirála (šroubovice). Je tvořena řetězci chemických částic – nukleotidů, které jsou vzájemně propojeny vazbami na základě typu báze, která je obsažena v nukleotidu. Řetězec si představíme jako sled bází. Existuje šest druhů bází – adenin (A), guanin (G), thymin (T), cytosin (C) a uracil (U). V DNA se vyskytují pouze čtyři z nich a spojují se dle komplementarity bází – vazby se tvoří mezi A – T a C – G (uracil nahrazuje ve vazbách thymin u RNA – kyseliny ribonukleové).



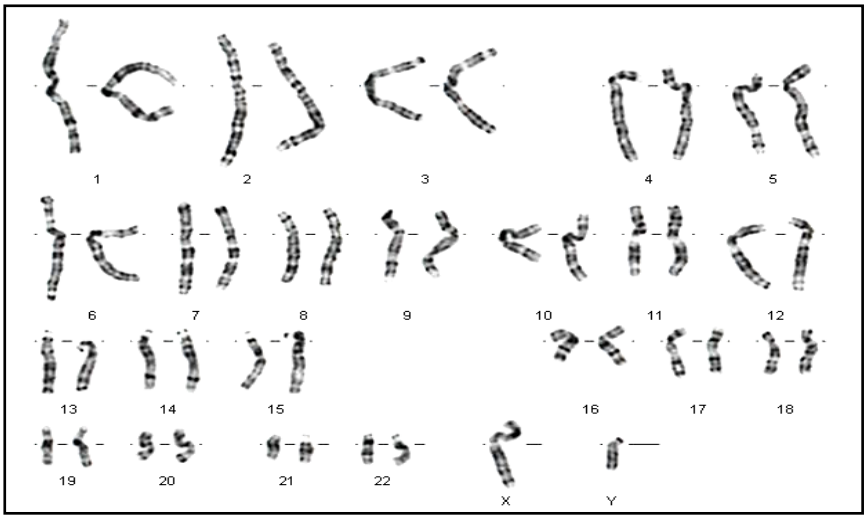
Obr. 1 DNA, upraveno dle [3]

Úsek molekuly DNA, který nese genetickou informaci pro určitý znak, se nazývá *gen*. Jeho umístění je dáno *lokusem* – místem na chromozomu v buňce daného organismu. Geny se dělí na dvě skupiny. *Geny velkého účinku*, kdy jeden gen má velký fenotypový účinek (na tvorbě kvalitativního znaku se podílí často jen jeden gen) a okolní prostředí má na projevení znaku malý význam. Druhou skupinou jsou *geny malého účinku*, jejichž fenotypový účinek je zanedbatelný. Na tvorbě kvantitativního/polygenního znaku, se podílí celý soubor genů a prostředí má velký vliv.

Kvalitativní znaky se vyskytují u jedinců v různých alternativách a nelze je číselně vyjádřit. Patří mezi ně např. barva oči, krevní skupina.

Kvantitativní, neboli polygenní *znaky* se objevují u jedinců v různých hodnotách. Můžeme je částečně číselně vyjádřit a jsou ovlivněny ze značné části prostředím. Např. tělesná výška, inteligence.

U diploidních organismů, mezi které patří i člověk, se nachází 2 sady chromozomů. Chromozomy tvoří páry homologních chromozomů, z nichž jeden chromozom pochází od matky a druhý od otce. Prvních 22 párů se nazývají autozomy, poslední dva chromozomy se nazývají gonozomy a netvoří homologní pár. Soubor všech chromozomů v jádře buňky se nazývá *karyotyp*.



Obr. 2 Karyotyp normálního muže, upraveno dle [4]

Diploidní buňku s kompletní sadou chromozomů (u člověka 2 x 23) nazýváme *zygota.*

Alelou se rozumí konkrétní forma genu. V diploidní buňce jsou obsaženy vždy dvě alely. Kombinace alel mohou být buď stejné, *homozygotní* (AA-dominantní homozygot, aa – recesivní homozygot) nebo různé, *heterozygotní* (Aa).

Souhrn všech forem genů (alel) organismu nazýváme genotypem a to, jak se znaky projeví navenek, *fenotypem.*

Specifická kombinace alel odkazující na různé lokusy, avšak děděná společně, představuje *haplotyp* (haploidní genotyp).

1.2 Přenos genetické informace

Přenos genetické informace mezi generacemi probíhá na základě rozmnožování. Lidé dědí genetickou informaci od svých rodičů při procesech zvaných mitóza a meióza. *Mitóza* je proces buněčného dělení, jehož výsledkem je vytvoření sesterských buněk, které nesou kopie kompletního souboru chromozomů. Při meióze dochází k produkci buněk s redukovaným počtem chromozomů. Zplozením potomka dochází k nové, odlišné (diverzitní) kombinaci alel, získaných od otce i od matky.

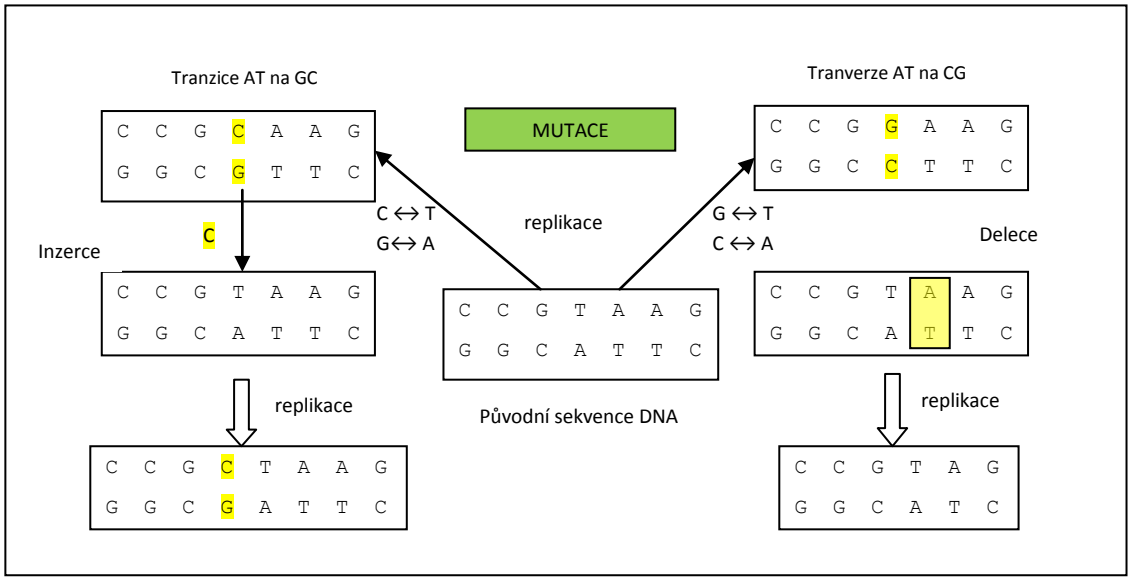
Při přenosu genetické informace může dojít k různým změnám, tzv. *mutacím*. Mutacemi vznikají nové formy genů, a tím vytvářejí větší genetickou variabilitu. Ke

změnám vyvolaným v důsledku faktorů vnějšího prostředí může dojít na úrovni struktury DNA, při transkripci nebo translaci.

Transkripce je proces přepisu genetické informace z vlákna DNA na mRNA (promediátorová kyselina RNA, která nese genetickou informaci, důležitou pro tvorbu bílkovin). *Translací* se rozumí překlad genetické informace z mRNA do pořadí aminokyselin v řetězci bílkoviny.

Podle rozsahu genetické informace, které mutace postihují, se může jednat o genové mutace (týkají se nukleotidové sekvence jednoho genu), chromozomové mutace (postihují DNA na úrovni chromozomů – změna struktury chromozomů) či genomové mutace (způsobují změnu počtu chromozomů).

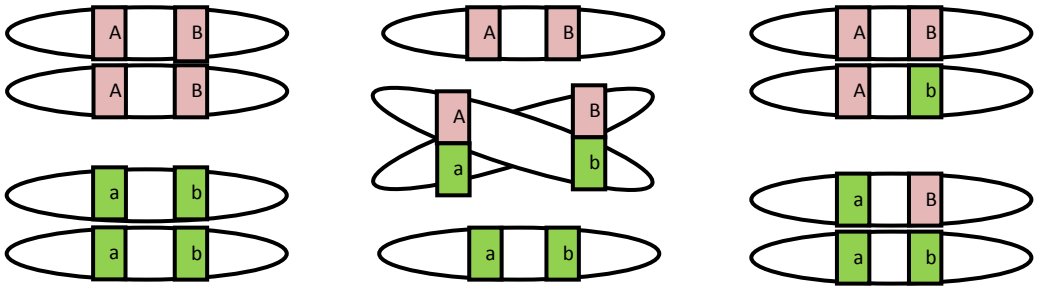
Mezi mutace na úrovni genů patří *inzerce* – vložení jednoho či více nukleotidů, *delece* – ztráta jednoho nebo více nukleotidů a *substituce* – záměna nukleotidů. Rozlišují se dva typy substitucí tranzice a tranzverze. *Tranzicí* se rozumí záměna *purinové báze* (báze A, G) za jinou purinovou bázi nebo pyrimidinové báze (báze T, C, U) za jinou pyrimidinovou bázi a *tranzverze* znamená záměnu purinu (báze A, G) za *pyrimidin* (báze T, C, U) nebo naopak.



Obr. 3 Genové mutace, upraveno dle [6]

Změny ve struktuře DNA mohou podmínit vznik strukturních změn chromozomů. Mezi takové změny patří právě polymorfismus chromozomů, jež je variantou některých chromozomů, ale neprojeví se na venek (nemají fenotypový efekt).

Další příčinou variability populace, kdy nedochází k vytvoření nových alel, ale k formování nových kombinací známých alel, je tzv. *crossing-over (rekombinace*). Rekombinace probíhá ve fázi meiózy, při tvorbě gamet (pohlavních buněk). Mezi homologickými chromozómy v těsné blízkosti proběhne vzájemná záměna úseků DNA, naruší se vazba genů. Pravděpodobnost rekombinace vrůstá se vzdáleností částí DNA na chromozomech.



Obr. 4 Rekombinace, upraveno dle [7]

Mírami genetické variability je polymorfismus a heterozygotnost populace. *Polymorfismus* je velké množství variant genů (alel) v jednom lokusu a je způsoben mnohými mutacemi a změnami v DNA. Udává podíl polymorfních lokusů v populaci. Přes 80% polymorfismů, jsou polymorfismy vzniklé na základě záměny jednoho nukleotidu – jednonukleotidové polymorfismy, značené *SNP* (single-nucleotid polymorfism). [9] *Heterozygotnost* populace je častěji používanou mírou genetické variability, protože je přesnější a spolehlivější. Stanoví se tím způsobem, že se určí četnosti heterozygotních jedinců v každém lokusu a vypočítá se průměr pro všechny lokusy. Je to tedy průměrná četnost heterozygotů v jednotlivých lokusech.

1.3 Hardyho-Weinbergova rovnováha (HWE – Hardy-Weinberg equilibrium)

HWE představuje nezávislost alel na daných lokusech mezi homologními chromozomy. Jinými slovy, znamená, že výskyt alely na jednom homologním chromozomu nezávisí na tom, jaká alela se vyskytuje na druhém homologním chromozomu. Je speciálním modelem pro předpověď genotypových četností v populaci. Vyjadřuje vztah mezi genotypovými a alelovými četnostmi.
Tento předpoklad (HWE) platí pouze za přesně daných podmínek:

- organizmy jsou diploidní,
- rozmnožování probíhá pohlavní cestou,
- v populaci se nevyskytuje migrace,
- lze zanedbat mutace,
- populace je panmiktická (páření je náhodné),
- populace je dostatečně velká.

Faktorem narušujícím platnost H.-W. zákona může být tzv. *inbreeding*. Při inbreedingu dochází k páření mezi příbuznými jedinci, tudíž je porušen předpoklad panmixie (náhodného páření). Snižuje se heterozygotnost populace (míra genetické variability).

Jsou-li dané podmínky splněny, lze v populaci stanovit genotypové četnosti pro gen se dvěma alelami.

Označíme-li p frekvenci alely A a q frekvenci alely q, pak platí, že frekvence tří fenotypů budou následující:

- pro AA: p^2,
- pro Aa: $2pq$,
- pro aa: q^2.

Pro frekvence alel vždy platí vztah

$$p + q = 1$$

a pro frekvence genotypů zase

$$p^2+2pq+q^2=1$$

Důležitým důsledkem HWE je, že četnosti alel v následující generaci zůstávají stejné (stálé) jako v generaci původní.

1.4 Vazebná nerovnováha

Vyjadřuje nenáhodné kombinace alel ve dvou či více lokusech v populaci a dochází k ní např. vlivem genetické vazby mezi alelami na jednom homologním chromozomu. Způsobuje, že změny v četnosti alel na jednom lokusu působí na změny v jiném lokusu. Vztahuje se i na asociaci více lokusů v těsné vazbě.

Vazebná analýza je přístup, který se užívá v rodinných studiích, které na rozdíl od populačních asociačních studií, zahrnují příbuzné jedince. Testuje spoluvýskyt markeru a fenotypu nemoci v rodině.

1.5 Studie případů a kontrol

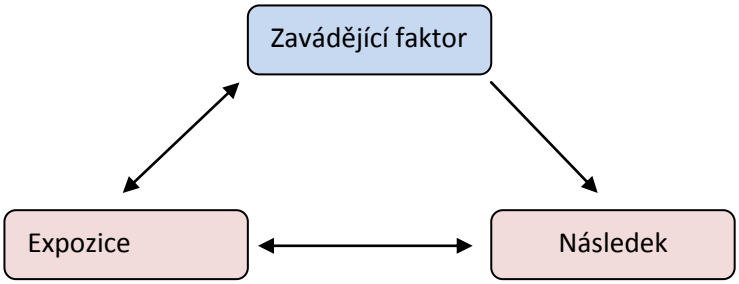
Populační asociační studie mívají nejčastěji podobu studie případů a kontrol (case-control study). Probíhá tak, že se nejprve stanoví skupina případů (jedinců se sledovanou nemocí) a tato skupina je pak porovnávána se skupinou jedinců, kteří nevykazují danou nemoc (kontroly). Navíc se bere v úvahu, zda byla v takto stanovených skupinách v minulosti expozice (vystavení) potenciálnímu rizikovému faktoru. Pokud je expozice vyšší mezi případy, pak tento faktor může být opravdu rizikovým. Pokud je tomu na opak, může zase jít o protektivní (ochranný) faktor.

U tohoto typu studie je velmi důležité precizně stanovit skupinu případů. Je třeba brát v úvahu to, že případy mají reprezentovat celou populaci.

Důležitou úlohou je také vybrat vhodné kontroly, tzn. takové jedince, kteří se co nejvíce budou podobat případům, až na to, že se u nich nevyskytl sledovaný znak. Výběr kontrol je komplikovaný, jelikož jedinci, kteří nevykazují sledovaný znak, mohou mít jiné zdravotní potíže, což může být zavádějící.

Zavádějící faktor (confounding factor) je definován jako proměnná, která souvisí s expozicí a je v přímém či nepřímém vztahu k následku. Např. zkoumáme-li asociaci mezi

expozicí – užívání alkoholu, a znakem – hladinou cholesterolu, matoucím faktorem může být kouření. Souvisí se sledovaným znakem a zároveň se více vyskytuje u jedinců požívajících ve velké míře alkohol. Matoucí faktor se neobjevuje jako mezikrok v příčinné posloupnosti (causal pathway) k nemoci. Neuvážení zavádějících faktorů může vést k chybnému závěru o sledované asociaci.



Obr. 5 Zavádějící faktor, upraveno dle [14]

Jednou z nejužívanějších metod k potlačení zavádějícího faktoru v případě kategoriálních dat, je *stratifikace*. Celý soubor rozdělíme na skupiny (strata), uvnitř nichž je potenciální matoucí faktor neměnný. Asociaci spočteme jednotlivě v každé skupině, pak vypočteme ukazatel asociace váženým průměrem ukazatelů jednotlivých skupin.

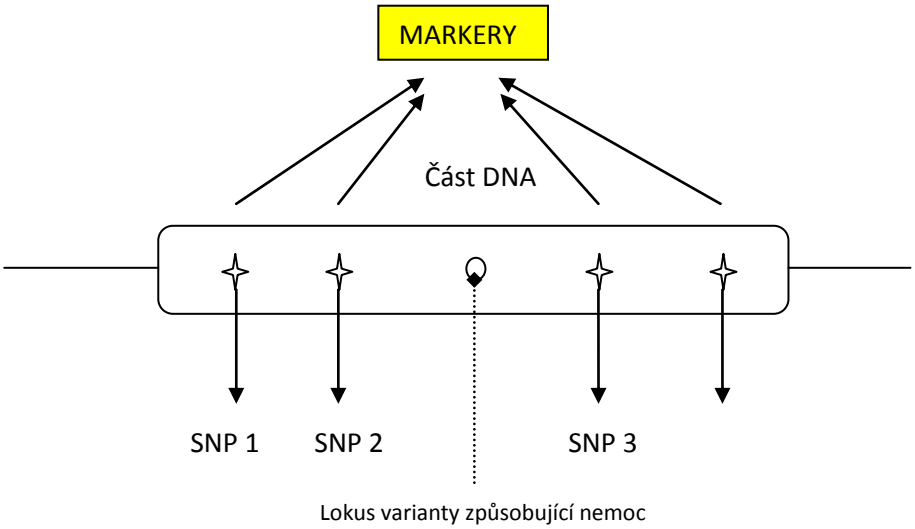
1.6 Typy výzkumů

Populační asociační studie mohou být rozděleny do těchto čtyř kategorií: kandidátní polymorfismus, kandidátní gen, detailní mapování, celogenomová studie.

Studie kandidátního polymorfismu jsou výzkumy asociací genotyp – znak. Polymorfismus je zde definován jako genetická varianta v jednom lokusu, která se vyskytuje minimálně v 1% populace. Cílem je testovat výskyt asociace a hypotézu, že daný polymorfismus (SNP) nebo více polymorfismů ovlivňují znak přímo (jsou funkční).

Ve *studiích kandidátních genů* se vyžaduje určení více SNP v rámci genu. Výběr SNP záleží na vazebné nerovnováze. Předpokladem těchto studií je, že tyto vybrané SNP zachycují informaci o genetické variabilitě genu, ačkoli nemusí přímo ovlivňovat znak

(nemoc), tj. nemusí být nutně funkční. Tyto SNP nejbližší k lokaci varianty podmiňující nemoc se nazývají *markery*, jsou asociovány s variantou, která nemoc působí.



Obr. 6 Markery, upraveno dle [1]

Cílem studií, nazývaných *detailní mapování*, je v genomu s vysokou mírou přesnosti určit umístění varianty způsobující nemoc. Znalost tohoto umístění umožňuje vyhnout se studiím založených na lokusech markerů.

Poslední typ studií, *celogenomová studie*, je zaměřena na zkoumání genetických variant v celém genomu - souboru všech struktur nesoucích genetickou informaci ve formě DNA (genom člověka obsahuje přes 20 000 genů [15]). Celogenomová studie je navržena tak, aby identifikovala asociace s pozorovanými znaky a určila tak známé sekvence DNA – markery pro výskyt nemoci. Pro celogenomové výzkumy je třeba větší počet SNP.

1.7 Zpracovávaná data

V průběhu dalšího textu budu užívat ke zpracování v systémech R (volně stažitelný na stránkách http://r-project.org/) a SAS (produkt společnosti SAS Institute, http://www.sas.com) veřejně dostupná data, která lze najít na webové stránce http://people.umass.edu/foulkes/asg/data/FMS_data.txt. Jedná se o funkční SNP asociované s velikostí a sílou svalů. Data byla shromážděna za účelem identifikace

rozhodujících činitelů velikosti a síly kosterního svalstva před a po cvičení (cvičení po dobu 12 týdnů). Soubor obsahuje data od 1397 jedinců (dobrovolníků z řad vysokoškolských studentů), celkem 225 SNP.

V souboru dat jsou genotypy SNP v genech a jsou zahrnuty i další proměnné (covariates) jako je období studijního roku (semestr) - term (1 - jaro, 2 - léto, 3- podzim), pohlaví (Gender), věk (age), rasa (race), %-ní změna síly dominantního deltového svalu před a po cvičení (DRM.CH) a %-ní změna síly nedominantního deltového svalu před a po cvičení (NDRM.CH).

2 Kontingenční tabulky pro binární znak

2.1 Poměr šancí (odds ratio)

Poměr šancí je dán jako poměr mezi šancemi výskytu jevu (onemocnění) v exponované populaci a neexponované populaci.

Šance výskytu (pravděpodobnost) jevu v exponované populaci je rovna

$$O(D^+|E^+) = \frac{P(D^+|E^+)}{1-P(D^+|E^+)}.$$

Podobně šance výskytu jevu v neexponované populaci

$$O(D^+|E^-) = \frac{P(D^+|E^-)}{1-P(D^+|E^-)}.$$

Poměr šancí definujeme jako

$$\mathrm{OR} = \frac{O(D^+|E^+)}{O(D^+|E^-)} = \frac{\frac{P(D^+|E^+)}{1-P(D^+|E^+)}}{\frac{P(D^+|E^-)}{1-P(D^+|E^-)}}, \qquad (1)$$

kde

D^+ značí výskyt nemoci,

D^- nepřítomnost nemoci ,

E^+ expozici faktoru,

E^- neexponovanou populaci.

Uvažujme, že budeme zkoumat asociaci mezi genotypem jedince a výskytem nemoci. Jak už jsem se zmiňovala, genotyp SNP může být dominantně homozygotní (AA), heterozygotní a recesivně homozygotní (aa). V případě, že se jedná o binární znak, můžeme data shrnout do kontingenční tabulky typu 2 x 3, kde n_{ij} znamená počet jedinců, pro i = 1,2 a j = 1,2,3.

		Genotyp			
		aa	Aa	AA	
Znak	+	n_{11}	n_{12}	n_{13}	$n_{1.}$
	-	n_{21}	n_{22}	n_{23}	$n_{2.}$
		$n_{.1}$	$n_{.2}$	$n_{.3}$	n

Počet jedinců, u kterých se projevila nemoc a vykazují genotyp aa

V genetice se nejčastěji počítá poměr šancí všech genotypů ve vztahu ke genotypu AA, tzn. $OR_{aa,AA}$ a $OR_{Aa,AA}$. Výpočet (dle vztahu (1)) bude vypadat takto

$$OR_{aa,AA} = \frac{(n_{11}/n_{.1}) / (n_{21}/n_{.1})}{(n_{13}/n_{.3}) / (n_{23}/n_{.3})}.$$

Po zkrácení dostaneme

$$ORaa,AA = \frac{n_{11}n_{23}}{n_{21}n_{13}}. \qquad (2)$$

Podobně pro poměr šancí genotypu Aa vzhledem k AA

$$OR_{Aa.AA} = \frac{(n_{12}/n_{.2}) / (n_{22}/n_{.2})}{(n_{13}/n_{.3}) / (n_{23}/n_{.3})} = \frac{n_{12}n_{23}}{n_{22}n_{13}}.$$

Výsledek nám říká, kolikrát vyšší je šance výskytu nemoci u exponované populace než u neexponované populace. Expozicí se zde přitom rozumí daný genotyp.

2.1.1 Poměr šancí v R

V systému R lze poměr šancí jednoduše spočítat pomocí funkce **`oddsratio()`** z balíčku **`epitools`** nebo přímým výpočtem z definice.

Pro data v tabulce 1 vypočítáme poměr šancí $OR_{aa,AA}$ přímo dle definice.

	aa	Aa	AA	celkem
znak +	332	164	215	711
znak -	230	262	225	717
celkem	562	426	440	1428

Tabulka 1

Data uspořádáme do matice pomocí funkce **`matrix()`**, kde volba **`nrow`** indikuje počet řádků matice a volba **`byrow`** zadává, že zadaná data se budou načítat do řádků, nikoli do sloupců, jak by tomu bylo, kdybychom tento příkaz vynechali. Funkcí **`colnames()`** zadáváme názvy sloupců, funkcí **`rownames()`** zase názvy řádků.

```
> data=matrix(c(332,164,215,230,262,225),nrow=2,byrow=T)
> colnames(data)=c("aa","Aa","AA")
> rownames(data)=c("znak+","znak-")
> data
       aa  Aa  AA
znak+ 332 164 215
znak- 230 262 225
```

Spočítáme poměr šancí genotypu aa vzhledem ke genotypu AA, tedy $OR_{aa,AA}$ dle vztahu (2).

```
> or.aaAA=(data[1,1]*data[2,3])/(data[2,1]*data[1,3])
> or.aaAA
[1] 1.510617
```

Výsledek vypovídá, že šance výskytu nemoci u exponované populace je přibližně 1,5-krát vyšší než u neexponované populace (expozice = daný genotyp).

2.1.2 Poměr šancí v systému SAS

Pro čtyřpolní kontingenční tabulky, které vzniknou vyloučením jedné z variant genotypu, např. kombinace alel Aa (heterozygot), lze vypočítat poměr šancí prostřednictvím úlohy **Table Analysis** v sekci **Describe** SAS Enterprise Guide (SAS EG).

Vhodné uspořádání dat pro úlohu **Table Analysis** je ve formátu kategoriálních proměnných pro znak (varianty např. 1 = přítomnost znaku, 2=nepřítomnost znaku) a pro genotyp (varianty 1 = aa , 2 = Aa, 3 = AA) a četnostní proměnné. Pro data z příkladu řešeného v systému R vypadá takováto datová množina např. takto

	znak	alela	pocet
1	1	1	332
2	1	2	164
3	1	3	215
4	2	1	230
5	2	2	262
6	2	3	225

Obr.7

a lze ji získat např. prostřednictvím kódu:

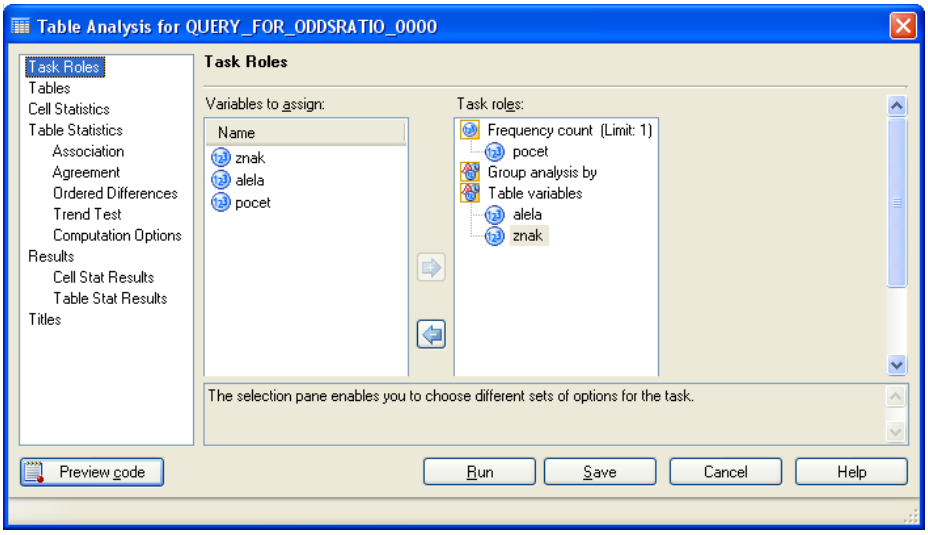
```
data oddsRatio;
input znak alela pocet;
datalines;

1 1 332
1 2 164
1 3 215
2 1 230
2 2 262
2 3 225
;
run;
```

Po úpravě (např. dotazem vytvořeným nástrojem **Query Builder** - volba **Filter and Query...** v menu **Data**), tj. vyloučením varianty „**Aa**“ v proměnné **alela** a úpravou formátu této proměnné získáme vstupní datovou množinu pro úlohu **Table Analysis** (viz obr. 8).

	znak	alela	pocet
1	1	aa	332
2	1	AA	215
3	2	aa	230
4	2	AA	225

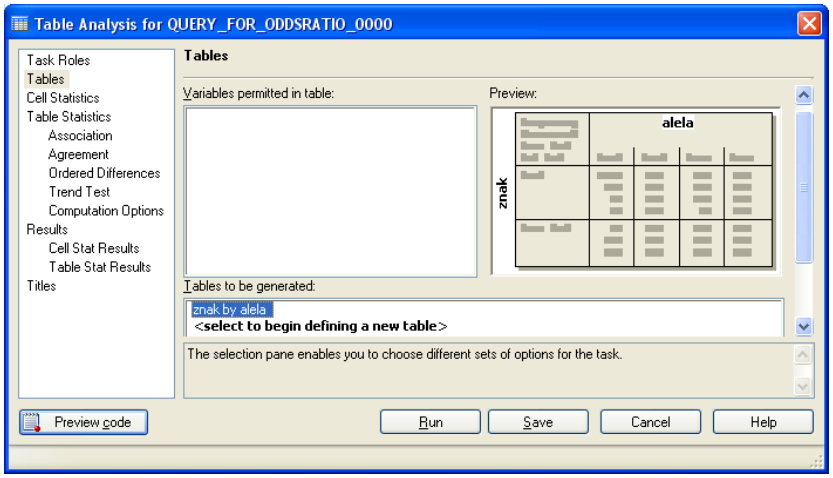
Obr.8

V úloze **Table Analysis** nastavíme role jednotlivých proměnných tak, jak je uvedeno na obr. 9.



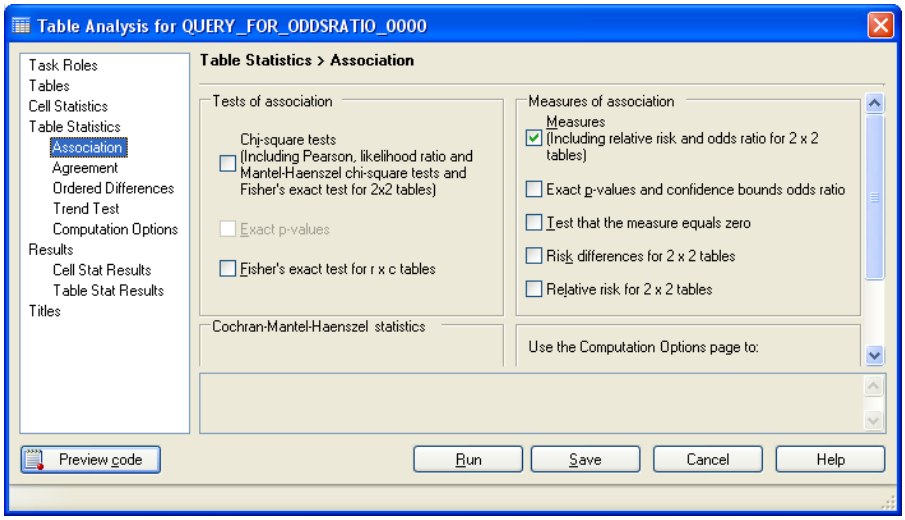
Obr. 9

Dále v sekci **Tables** vytvoříme tabulku pro analýzu kombinací sloupcové proměnné **`alela`** a řádkové **`znak`** (viz obr. 10).



Obr.10

V sekci **Table Statistics** a podsekci **Association** nastavíme výpočet poměru šancí zaškrtnutím první volby v části **Measures of association** (obr. 11).



Obr. 11

Ve výstupu úlohy potom vidíme vlastní kontingenční tabulky pro vybrané varianty genotypu a vypočtený poměr šancí $OR_{aa,AA}$ (obr. 12).

Frequency Col Pct	Table of znak by alela		
	alela		
znak	aa	AA	Total
1	332 59.07	215 48.86	547
2	230 40.93	225 51.14	455
Total	562	440	1002

Estimates of the Relative Risk (Row1/Row2)			
Type of Study	Value	95% Confidence Limits	
Case-Control (Odds Ratio)	1.5106	1.1748	1.9424
Cohort (Col1 Risk)	1.2007	1.0722	1.3446
Cohort (Col2 Risk)	0.7948	0.6913	0.9139

Obr. 12

2.2 Pearsonův χ^2 – test

Pomocí Pearsonova χ^2 – testu nezávislosti testujeme nulovou hypotézu, že náhodná veličina (znak) a expozice (genotyp) jsou nezávislé, tj.

$$H_0: OR = 1.$$

Důležitým předpokladem pro použití tohoto testu je, že máme k dispozici dostatečně velký výběr z populace a nejmenší četnost v kontingenční tabulce splňuje

podmínku $n_{ij} > 5$, i=1,..,r, j=1,..,s, kde i a j jsou varianty statistických znaků uspořádaných do kontingenční tabulky. Pokud není tento předpoklad splněn, není užití Pearsonova testu vhodné a upřednostníme místo něj tzv. Fisherův exaktní test.

Při aplikaci Pearsonova χ^2 testu postupujeme tak, že odvodíme očekávané četnosti E_{ij}, i=1,..,r, j=1,..,s, za předpokladu nezávislosti genotypu a znaku. Dostaneme je tedy tak, že vynásobíme pravděpodobnosti genotypu a znaku a podělíme rozsahem výběru, tj.

$$E_{ij} = \frac{n_{i.} n_{.j}}{n},$$

kde i=1,..r, j=1,..,s a n je celkový počet pozorování (jedinců), tj. rozsah výběru.

Pozorované četnosti označíme jako náhodnou proměnnou O_{ij}, za kterou potom v konkrétním případě dosazujeme hodnoty četností n_{ij}.

Testová statistika Pearsonova χ^2 testu je obecně dána předpisem

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \sim \chi^2_{(r-1)(s-1)}$$

Statistika má asymptoticky χ^2 rozdělení o (r-1)(s-1) stupních volnosti, kde *r* představuje počet řádků a *s* počet sloupců kontingenční tabulky. Pokud vypočtená hodnota testovací statistiky splňuje podmínku

$$\chi^2 \geq \chi^2_{(r-1)(s-1)}(1-\alpha),$$

kde $\chi^2_{(r-1)(s-1)}(1-\alpha)$ je (1-α)-kvantil rozdělení $\chi^2_{(r-1)(s-1)}$, hypotézu H_0 zamítáme na hladině významnosti α (obvykle 0,05 nebo 0,01).

Místo porovnání hodnoty testovacího kriteria s kritickou hodnotou (kvantilem) můžeme použít též p-hodnoty, která bývá využívána zejmena při práci se softwary.
P-hodnota je pravděpodobnost, s jakou testovací statistika nabývá hodnot více svědčících proti testované hypotéze, udává mezní hladinu významnosti, při které bychom hypotézu ještě zamítali. Pokud je p-hodnota menší než stanovená hladina významnosti α, hypotéza H_0 se zamítá.

2.2.1 Pearsonův χ^2 test v systému R

Předpokládejme, že chceme zjistit, zda existuje asociace mezi některým SNP v genu esr1 a BMI (body mass indexem) větším jako 25 na základě dat, která máme k dispozici.

Začneme specifikací umístění dat. Objekt **fms** bude proměnná uchovávající URL odkaz na umístění souboru s daty na internetu. Stejně tak lze použít i odkaz na umístění souboru s daty na disku počítače, tzv. cestu, ať už relativní (začíná v aktuálním pracovním adresáři) nebo absolutní (začíná označením diskové jednotky počítače).

```
> fms = "http://people.umass.edu/foulkes/asg/data/FMS_data.txt"
```

Následně použijeme k vložení dat do R funkci **read.delim()**, která slouží k načítání dat oddělených specifikovaným oddělovačem, nejčastěji čárkou – soubory ve formátu CSV (Comma Separated Value). Specifikací **header = T** určíme, že první řádek souboru bude obsahovat názvy proměnných, tzv. hlavičku. Volba **sep="\t"** znamená, že vkládáme data, kde oddělovačem je tabulátor.

```
> fms = read.delim(file=fms, header=T, sep="\t")
```

Určíme názvy všech SNP genu esr1prostřednictvím funkce **names()**, která vrací názvy proměnných v objektu a funkce **substr()**, která nalezne podřetězec v zadaném řetězci.

```
> NamesEsr1Snps = names(fms)[substr(names(fms),1,4)=="esr1"]
> NamesEsr1Snps
[1] "esr1_rs1801132" "esr1_rs1042717" "esr1_rs2228480" "esr1_rs2077647"
[5] "esr1_rs9340799" "esr1_rs2234693"
```

Zápis **substr(names(fms),1,4)=="esr1"** znamená porovnání („**==**“..rovnost) prvních čtyř (1.. první znak podřetězce=1, 4..počet znaků podřetězce=4) znaků každého řetězce z vektoru názvů, který vrací funkce **names()** aplikovaná na objekt **fms**, s řetězcem „esr1“. Výsledkem tohoto porovnání je vektor logických hodnot (**TRUE**..pravda, **FALSE**..nepravda), který poté slouží k výběru vyhovujících prvků

z řetězce názvů proměnných. V našem případě vyhovuje této podmínce celkem šest proměnných, jak je vidět, když zobrazíme hodnotu proměnné **NamesEsr1Snps.**

Genotypovou matici nyní můžeme definovat výběrem sloupců (druhá dimenze objektu **fms**, proto je podmínka výběru uvedena až za znakem",", vynechání výběru v první dimenzi znamená, že vybíráme všechny řádky objektu **fms**), které odpovídají názvům esr1 SNP.

```
> fmsEsr1 = fms[,is.element(names(fms),NamesEsr1Snps)]
```

Funkce **is.element()** je jednou z tzv. informačních funkcí, které vracejí logickou hodnotu pro každý prvek objektu, který je zadán jako jejich první argument. V našem případě prohledáváme prvky vektoru názvů proměnných v objektu **fms** a porovnáváme je s názvy uloženými v proměnné **NamesEsr1Snps**, která je uvedena jako druhý parametr.

Jako znak nadefinujeme skutečnost, kdy BMI (body mass index, proměnná **pre.BMI**) > 25. Funkce **as.numeric()** patří mezi konverzní funkce, které slouží ke konverzi formátů objektů. V tomto případě konvertujeme logickou hodnotu **TRUE** nebo **FALSE** na číslo **1** nebo **0**.

```
> Trait = as.numeric(pre.BMI>25)
```

Následně nadefinujeme funkci **newFunction**, která vypíše tzv. p-hodnoty generované prostřednictvím χ^2-testu, tedy funkce **chisq.test()**.

```
> newFunction = function(Geno){
+ ObsTab = table(Trait,Geno)
+ return(chisq.test(ObsTab)$p.value)
+ }
```

Funkce **table()** vytváří tabulku četností. V našem případě kontingenční tabulku četností znaku **Trait** (BMI > 25) a znaku, který bude do funkce předán jako parametr **Geno**. Tyto proměnné by měly mít stejnou délku (neznamená to však stejný počet variant

znaku). Zápis **chisq.test(ObsTab)$p.value)** znamená, že z tzv. návratové hodnoty funkce **chisq.test()** chceme pouze tu část, která je označena jako **p.value** (odpovídá p-hodnotě χ^2 testu aplikovaného na konkrétní data).

Vytvořenou funkci použijeme na sloupce, jejichž název je uložen v proměnné **fmsEsr1**. Využijeme k tomu funkci **apply()**, která aplikuje funkci **newFunction** (3. parametr) na sloupce (**2**..sloupce, **1**..řádky) objektu **fmsEsr1** (1. parametr).

```
> apply(fmsEsr1,2,newFunction)
esr1_rs1801132 esr1_rs1042717 esr1_rs2228480 esr1_rs2077647 esr1_rs9340799
0.4440720      0.0264659      0.1849870      0.1802880      0.1606800
esr1_rs2234693
0.1675418
```

Na základě výsledků, vidíme, že je pravděpodobná asociace mezi druhým SNP (**esr1_rs1042717**) a BMI. U této proměnné je p-hodnota χ^2 testu nezávislosti menší než stanovená hodnota hladiny významnosti 0,05).

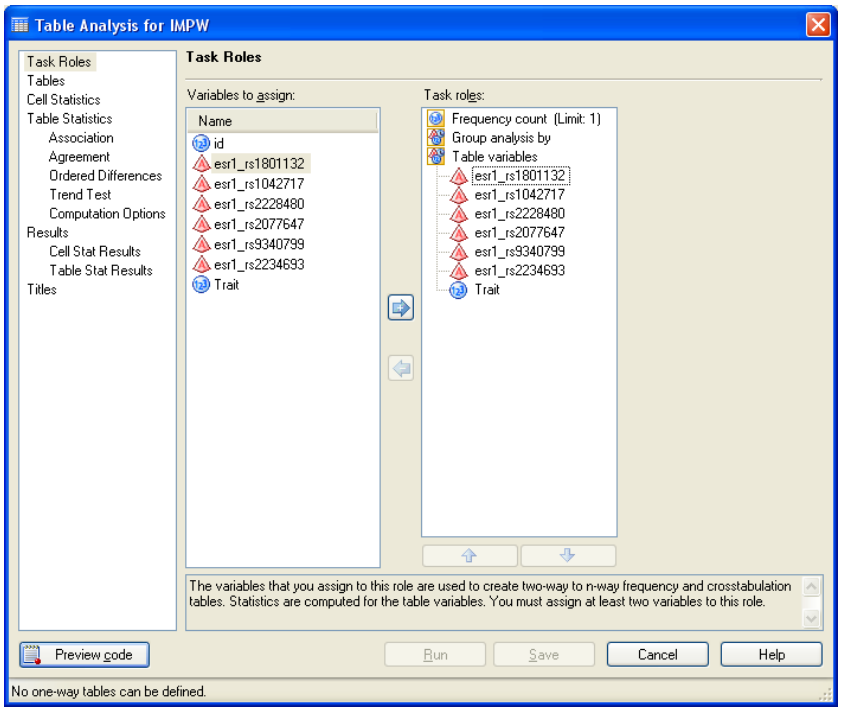
2.2.2 Pearsonův χ^2 test v systému SAS

Předpokládejme, že máme vytvořenou datovou množinu, která obsahuje pouze proměnné id, **esr1_rs1801132**, **esr1_rs1042717**, **esr1_rs2228480**, **esr1_rs2077647**, **esr1_rs9340799**, **esr1_rs2234693** a proměnnou **Trait**, která vznikla na základě podmínky **pre.BMI>25** (obr. 13).

	id	esr1_rs1801132	esr1_rs1042717	esr1_rs2228480	esr1_rs2077647	esr1_rs9340799	esr1_rs2234693	Trait
1	1	CG	GG	GG	GG	GG	CC	1
2	2	CG	GA	GA	AA	AA	TT	1
3	3	CG	GG	GG	AG	AG	TC	0
4	4	CC	GG	GA	AG	AA	TC	1
5	5	CG	GG	GG	AA	AA	TT	0
6	6	CC	GG	GG	GG	AG	TC	0
7	7	CC	GG	GG	AA	AA	TT	0
8	8	CG	GG	GG		AA	TT	.
9	9	CG	GG	GG	AA	AA	TT	0
10	10	CC	GG	GA	AA	AA	TT	1
11	11	CC	GG	GG	AG	AG	TC	.
12	12	CC	GG	GA	AA	AA	TT	0
13	13	CG	GG	GG	AG	AG	TC	0
14	14	CC	GG	GG	GG	GG	CC	0
15	15							.
16	16	CG	GA	GA	AA	AA	TC	0
17	17	CC	GG	GG	GG	GG	CC	.
18	18	CG	GG	GG	AA	AA	TC	.

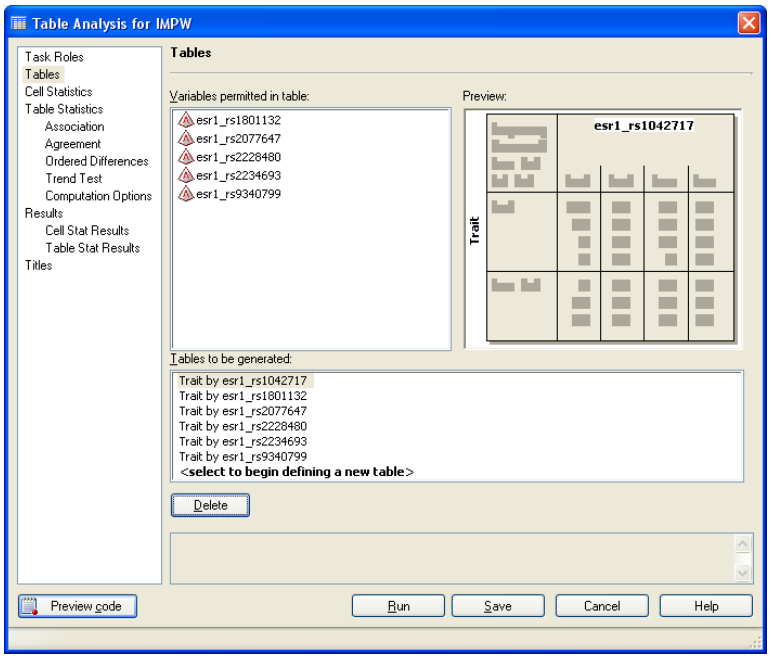
Obr. 13

Pearsonův χ^2 test můžeme opět realizovat prostřednictvím úlohy **Table analysis** jako u poměru šancí. Nastavení rolí provedeme dle obr. 14.



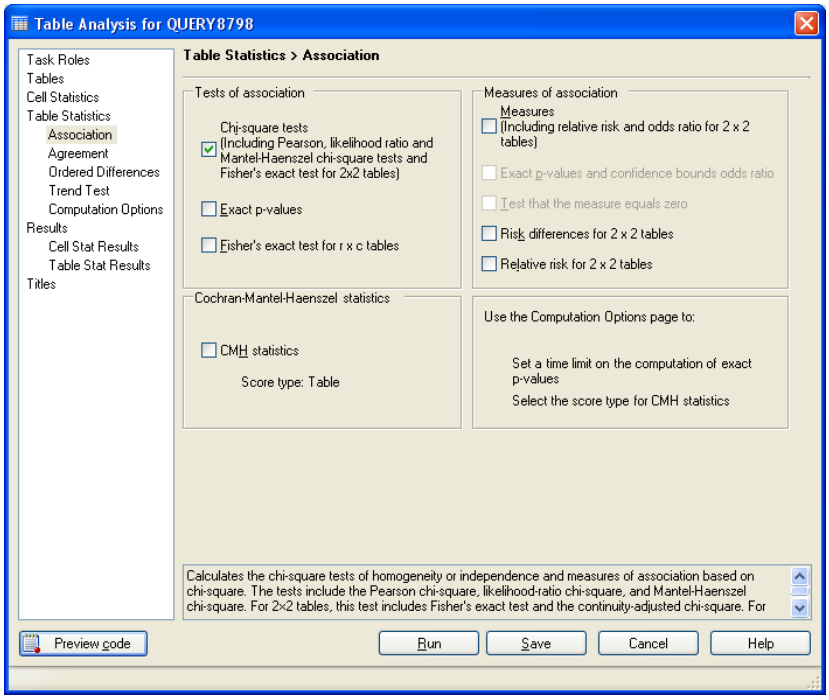
Obr. 14

V sekci **Tables** můžeme nadefinovat kontingenční tabulky pro všechny vybrané SNP (obr. 15).



Obr. 15

V sekci **Table Statistics**, podsekci **Association**, části **Test of association** zaškrtneme první volbu **Chi-square tests** (obr. 16).



Obr. 16

Pro SNP např. **`esr1_rs1042717`**, u kterého jsme zpracováním v R zjistili možnou závislost, pak ve výsledcích najdeme kontingenční tabulku i p-hodnotu Pearsonova testu (viz obr. 17) a zjistíme, že p-hodnoty v obou softwarech vyjdou podobně, tedy je možná závislost mezi tímto SNP a znakem (BMI > 25).

Frequency
Col Pct

Table of Trait by esrl_rs1042717

Trait(Trait)	esrl_rs1042717(esrl_rs1042717) AA	GA	GG	Total
0	30 50.00	246 65.43	380 67.38	656
1	30 50.00	130 34.57	184 32.62	344
Total	60	376	564	1000

Frequency Missing = 398

Statistics for Table of Trait by esrl_rs1042717

Statistic	DF	Value	Prob
Chi-Square	2	7.2638	0.0265
Likelihood Ratio Chi-Square	2	6.9448	0.0310
Mantel-Haenszel Chi-Square	1	4.4921	0.0341
Phi Coefficient		0.0852	
Contingency Coefficient		0.0849	
Cramer's V		0.0852	

Obr. 17

2.3 Fisherův exaktní test

Jak už bylo řečeno, užití Fisherova exaktního (faktoriálového) testu je vhodnější, pokud máme malý rozsah výběru n nebo je nejmenší četnost v kontingenční tabulce menší než 5.

Při aplikaci Fisherova exaktního testu postupujeme tak, že nejprve vypíšeme všechny možnosti kontingenčních tabulek, při daných marginálních četnostech $n_{1.}, n_{2.}, n_{.1}, n_{.2}$.

Následně u nich vypočteme hodnoty pravděpodobností P

$$P = \frac{P(n_{11},n_{12},n_{21},n_{22})}{R} = \frac{n_{1.}!n_{2.}!n_{.1}!n_{.2}!}{n!n_{11}!n_{12}!n_{21}!n_{22}!}$$

A hodnoty logaritmických interakcí δ, dle vztahu

$$\delta = \ln\frac{p_{11}p_{22}}{p_{12}p_{21}},$$

přičemž součet všech pravděpodobností P musí dát 1.

Provedení testu nezávislosti hypotézy

$$H_0: \delta = 0 \qquad (3)$$

při oboustranné alternativě H_A: δ≠ 0 probíhá tak, že sečteme pravděpodobnosti P všech tabulek, u nichž absolutní hodnota logaritmické interakce je větší nebo rovna absolutní hodnotě logaritmické interakce výchozí kontingenční tabulky.
Pokud je výsledný součet menší jak stanovená hladina významnosti α, zamítáme H_0 (3), tudíž není prokázána asociace mezi sledovanými znaky.

2.3.1 Fisherův exaktní test v systému R

Budeme nás opět zajímat asociace mezi SNP esr1 genu a BMI > 25. Vytvoříme funkci, která vypočítá p-hodnoty Fisherova exaktního testu asociace mezi každým SNP a znakem (BMI). Fisherův exaktní test provádí funkce **`fisher.test()`**.

```
> newFunction = function(Geno){
+ ObsTab = table(Trait,Geno)
+ return(fisher.test(ObsTab)$p.value)
```

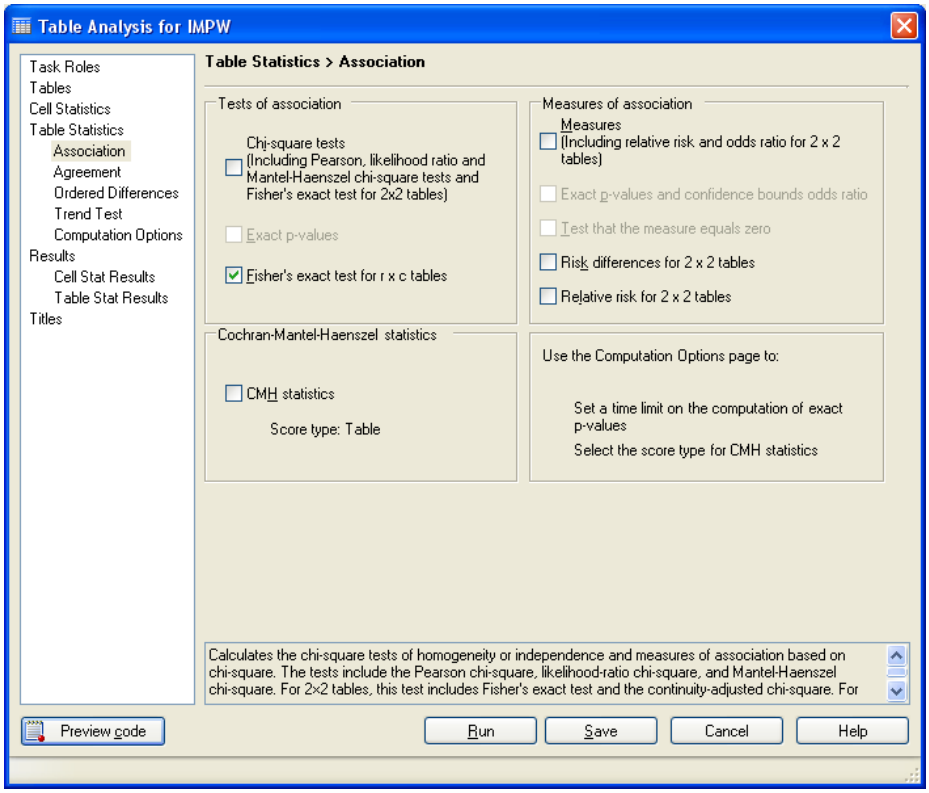
Tuto novou funkci aplikujeme na sloupce esr1, tj. sloupce objektu **`fmsEsr1`**, opět pomocí funkce **`apply()`**.

```
> apply(fmsEsr1,2,newFunction)
esr1_rs1801132 esr1_rs1042717 esr1_rs2228480 esr1_rs2077647 esr1_rs9340799
    0.46053113     0.02940733     0.18684765     0.17622428     0.15896064
esr1_rs2234693
    0.16945636
```

P-hodnoty Fisherova exaktního testu jsou srovnatelné s p-hodnotami χ^2 testu. Stejně jako u χ^2 testu výsledek naznačuje asociaci mezi druhým SNP a BMI > 25 (zde je p-hodnota menší jak $\alpha = 0{,}05$).

2.3.2 Fisherův exaktní test v systému SAS

Analogicky jako pro Pearsonův test nezávislosti, využijeme v prostředí SAS EG úlohu **Table Analysis**. Při zadávání postupujeme stejně jako u χ^2 testu, pouze v sekci **Table Statistics**, podsekci **Association** zaškrtneme volbu **Fisher's exact test for r x c tables** (obr. 18).



Obr.18

Pro SNP **`esr1_rs1042717`**, u kterého výpočtem v systému R vyšla p-hodnota značící závislost mezi tímto SNP a znakem, nalezneme ve výsledcích podobnou p-hodnotu Fisherova exaktního testu 0.0294 (obr. 19).

Fisher's Exact Test	
Table Probability (P)	1.857E-04
Pr <= P	0.0294

Obr. 19

2.4 Korelace

Termín korelace je často používán ve smyslu závislosti mezi dvěma proměnnými. Pro výpočet korelace se užívá Pearsonův a Spearmanův korelační koeficient.

Pearsonův korelační koeficient vyjadřuje lineární závislost dvou náhodných spojitých veličin X,Y. Může nabývat hodnot z intervalu $\langle -1, 1 \rangle$. Pokud jsou náhodné veličiny X,Y normálně rozdělené a korelační koeficient rovná 0 (jejich kovariance je nulová), není mezi X,Y lineární závislost. Musíme si však uvědomit, že závislost může existovat i jiná než lineární.

Pearsonův korelační koeficient mezi náhodnou veličinou X a Y se vypočítá dle vztahu

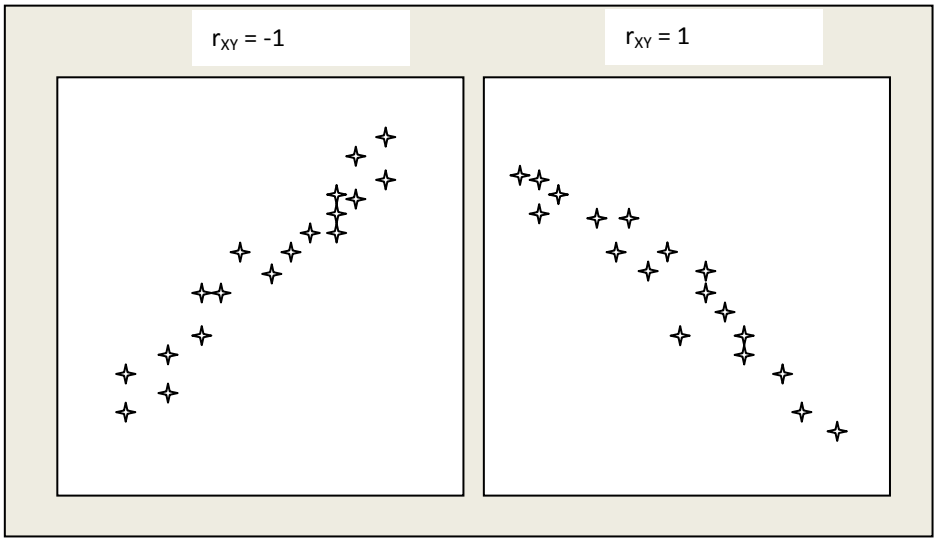
$$\rho\, XY = \frac{S_{xy}}{\sqrt{S_x^2 S_y^2}} = \frac{E(XY) - E(X)E(Y)}{\sqrt{(E(X^2) - E^2(X))(E(Y^2) - E^2(Y))}}$$

Pokud sledujeme dva znaky, které nabývají dvou různých variant, např. výskyt nemoci vzhledem ke genotypům AA,aa a Aa, můžeme možnosti shrnout do kontingenční tabulky typu 2x2

	AA,aa	Aa	
Nemoc se vyskytuje	n_{11}	n_{12}	$n_{1.}$
Nemoc se nevyskytuje	n_{21}	n_{22}	$n_{2.}$
	$n_{.1}$	$n_{.2}$	n

a počítáme tzv. výběrový korelační koeficient takto

$$r_{XY} = \frac{n_{11}n_{22} - n_{12}n_{21}}{\sqrt{n_{1.}n_{2.}n_{.1}n_{.2}}}$$



Obr. 20 Lineárně závislá data, upraveno dle [15]

Spearmanův korelační koeficient je neparametrická metoda zjištění korelace, založená na pořadích jedinců uspořádaných dle velikosti, vzhledem ke dvěma sledovaným veličinám.

Užívá se v případě, že je narušen předpoklad normality nebo nemůžeme-li hodnoty náhodných veličin přesně zjistit, ale máme k dispozici pořadí veličin Q (pořadí jedinců dle první veličiny) a R (pořadí jedinců dle druhé veličiny). Jsou-li si tato pořadí podobná, značí to závislost mezi náhodnými veličinami.

Spearmanův korelační koeficient vypočítáme za pomocí diferencí $d_i = Q_i - R_i$, $i = 1,..,n$ podle vztahu

$$r_s = 1 - \frac{6}{n(n^2-1)} \sum_{i=1}^{n} d_i^2$$

Spearmanův korelační koeficient rovněž nabývá hodnot mezi -1 a 1 a pokud je roven 0, naznačuje, že mezi sledovanými veličinami není lineární závislost.

2.5 Cochran-Armitage test (test trendu)

Test trendu užíváme k odhalení lineárního trendu. Předpokladem pro užití Cochran – Armitage testu je, že veličiny v kontingenční tabulce jsou ordinální (založené na uspořádaných kategorií, např. expozice nízká, střední, vysoká). Pokud je veličina spojitá, můžeme test trendu použít, pokud ji převedeme na ordinální, tzn. vhodně uspořádáme do kategorií. Musíme však brát v úvahu, že počet kategorií by neměl být příliš velký, jelikož k testování lineárního trendu se užívá statistika χ^2 testu, a mohlo by se stát, že v kategoriích klesne obsazení tak, že bude porušen předpoklad minimální četnosti $E_{ij} > 5$, $i=1,..,r$, $j=1,..,s$.

Pro jednoduchost se omezíme na případ s obecně k ordinálně uspořádanými kategoriemi expozice a dvěma kategoriemi následku.

Uvažujme, že budeme zkoumat vztah mezi expozicí (genotypem) a výskytem nemoci. Genotypy můžeme převést do tří ($k = 3$) uspořádaných kategorií na základě počtu alely A, tj.

AA – 2,

Aa – 1,

Aa – 0.

Hodnoty četností genotypu a znaku (např. nemoci) shrneme do kontingenční tabulky pro test trendu.

Kategorie	Skór	Nemoc + (případy)	Nemoc – (kontroly)	Celkem	Podíl jedinců se znakem
0	x_0	y_0	$n_0 - y_0$	n_0	y_0/n_0
1	x_1	y_1	$n_1 - y_1$	n_1	y_1/n_1
2	x_2	y_2	$n_2 - y_2$	n_2	y_2/n_2

V každé kategorii se předpokládá binomické rozdělení počtu následků (přítomnosti nemoci).

Pravděpodobnost že v v i-té kategorii nastane znak (vyskytne se nemoc) je

$$\pi_i = \alpha + \beta x_i,$$

kde i = 0,1,2 a x_i jsou skóry.

Skóry kvantifikují postavení kategorie v kontextu ostatních. Mohou jimi být středy tříd (např. medián) nebo je volíme symetricky kolem nuly (-1,0,1), popřípadě jako pořadí (1, 2, 3).

Testem lineárního trendu testujeme nulovou hypotézu, že směrnice přímky je rovna 0 (neexistuje lineární závislost)

$$H_0: \beta = 0. \qquad (4)$$

Postupujeme podobně jako u lineární regrese a dostaneme odhady pravděpodobností

$$\hat{\pi}_i = a + bx_i .$$

Odhad směrnice β vypočteme takto

$$b = \frac{\sum_{i=0}^{2} n_i(p_i - \bar{p})(x_i - \bar{x})}{\sum_{i=0}^{2} n_i(x_i - \bar{x})^2},$$

kde $\bar{p}$ je celkový podíl případů $\bar{p} = \frac{\sum_{i=0}^{2} y_i}{n}$

a $\bar{x}$ vypočítáme dle vztahu $\bar{x} = \frac{\sum_{i=1}^{k} x_i n_i}{n}$,

přičemž n je celkový počet jedinců, zahrnutých do studie.

Absolutní člen *a* dopočteme dle vztahu $a = \bar{p} - b\bar{x}$.

Nyní můžeme porovnat p_i s odhadem $\hat{\pi}_i$. Pokud jsou si hodnoty v kategoriích blízké, můžeme předpokládat lineární závislost π_i na hodnotách $x_{i.}$. Náš předpoklad ověříme testovou statistikou testu dobré shody

$$\chi^2_{\text{linearita}} = \frac{\sum_{i=0}^{2} n_i (p_i - \hat{\pi}_i)^2}{\bar{p}(1-\bar{p})} \sim \chi^2_{k-2}.$$

Za platnosti našeho předpokladu existuje lineární asociace mezi π_i a x_i, i = 1,…,k.

Nyní můžeme přejít k testovací statistice Cochranova-Armitageova testu a ověřit, zda lineární trend je statisticky významný, vrátíme se tedy k hypotéze (4).
Za předpokladu, že platí nulová hypotéza (4), pak platí

$$\chi^2_{\text{směrnice}} = \frac{b^2 \sum_{i=0}^{2} n_i (x_i - \bar{x})^2}{\bar{p}(1-\bar{p})} \sim \chi^2_1.$$

Je-li hodnota testovací statistiky rovna nebo větší jak $\chi^2_1(1-\alpha)$, zamítáme nulovou hypotézu ve prospěch hypotézy alternativní, že lineární trend je významný.

2.5.1 C-A test v systému R

C-A test se aplikuje pomocí funkce **`independence_test()`** z balíčku **`coin`**.

V příkladu se zaměříme na asociaci mezi **`esr1_rs1042717`** SNP a BMI (binární znak BMI ≤ 25, BMI > 25).

Začneme nainstalováním potřebného balíčku **coin**. Můžeme postupovat prostřednictvím nabídky **Packages**, a volby **Install Package(s)...** spolu s volbou **Load package...** nebo použít funkci **install.packages()** a funkci **library()**. První volbu (resp. funkci) použijeme jen jednou – slouží k instalaci balíčku na náš počítač, druhou pokaždé, když chceme nějakou funkci nebo data z konkrétního balíčku použít. Druhý způsob je výhodnější, pokud chceme provádět C-A test v rámci nějakého komplexnějšího skriptu.

```
> install.packages("coin")
> library(coin)
```

Nyní nadefinujeme genotyp a znak (není nutné vyřadit pozorování s chybějícími hodnotami genotypů).

```
> attach(fms)
> Geno = esr1_rs1042717
> Trait = as.numeric(pre.BMI>25)
```

Funkce **attach(fms)** umožní „vstup“ dovnitř objektu **fms**, tj. nebude nutné pro přístup k proměnným tohoto objektu používat předponu „**fms$**“. Toto nastavení rušíme pomocí funkce **detach()**.

V následujícím kroku uspořádáme genotypy do ordinálních kategorií 0 (aa), 1(Aa), 2 (AA) dle počtu alel A a aplikujeme funkci **independence_test()**. Zvolíme-li **testat = "quad"**, bude aplikován C-A test. Skóry specifikují vztah mezi kategoriemi genotypů. Funkce **ordered()** (analogie funkce **as.ordered()**) zajistí, že jednotlivé úrovně faktorové (kategoriální) proměnné budou chápány jako ordinální (uspořádané). Zápis **Trait~GenoOrd** znamená předpis (formuli) používaný pro vyjádření závislosti proměnných. Funkce **list()** vytváří objekt typu list, tj. seznam (heterogenní struktura, v našem případě složená z jediného prvku, a to vektoru označeného jako **GenoOrd**). Funkce **c()** vytváří vektor (homogenní datovou strukturu – prvky musí být stejného typu) o složkách zadaných jako parametry této funkce.

```
> GenoOrd = ordered(Geno)
```

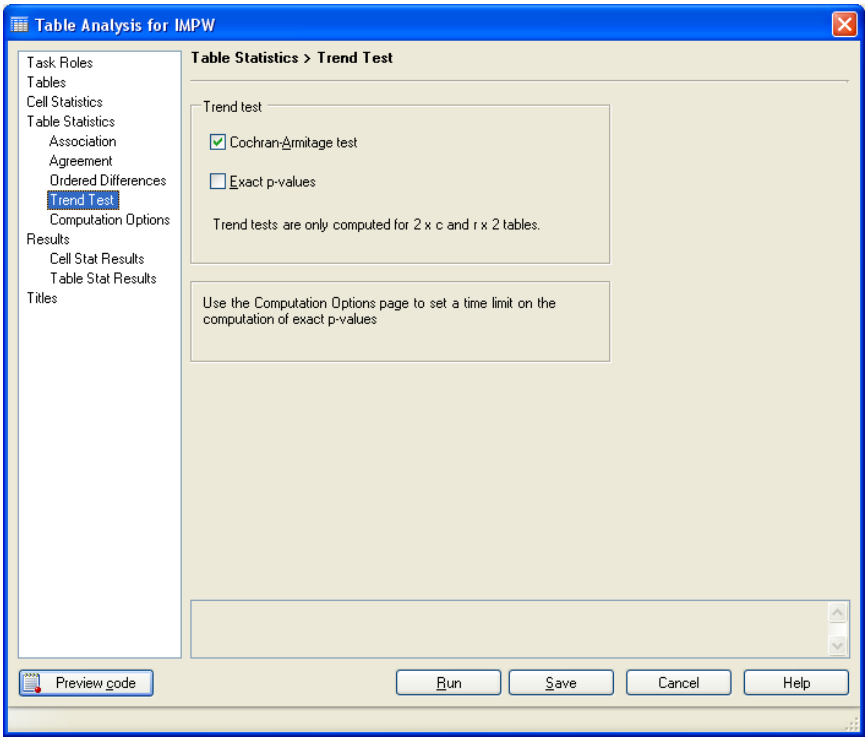
```
> independence_test(Trait~GenoOrd,teststat="quad",
scores=list(GenoOrd=c(0,1,2)))


 Asymptotic General Independence Test

data:  Trait by GenoOrd (AA < GA < GG)
chi-squared = 4.4921, df = 1, p-value = 0.03405
```

V případě užití testu trendu je p-hodnota větší než při užití χ^2 testu. Můžeme vidět, že poměr jedinců s BMI větším než 25 neklesá lineárně s genotypem tohoto SNP. Pokud si zobrazíme do matice počty jedinců s jednotlivými genotypy a binární znak (BMI ≤ 25 (0), BMI > 25 (1)) a spočítáme poměry jedinců s daným genotypem, s BMI > 25 (1), můžeme vidět, že poměr jedinců s BMI větším než 25 neklesá lineárně s genotypem tohoto SNP.

```
> data=matrix(c(30,246,380,30,130,184,60,376,564),nrow=3,byrow=T)
> colnames(data)=c("AA","GA","GG")
> rownames(data)=c("0","1","celkem")
> data
       AA   GA    GG
0      30   246   380
1      30   30    184
celkem 60   376   564
> pomery=c(data[2,1]/data[3,1],data[2,2]/data[3,2],data[2,3]/data[3,3])
> pomery
[1] 0.5000000 0.3457447 0.3262411
```

2.5.2 C-A test v systému SAS

Ve stejné úloze **Table Analysis**, ve které jsme již řešili nezávislost proměnných uspořádaných do kontingenční tabulky pomocí Pearsonova χ2 testu a Fisherova exaktního testu, lze nastavit i výpočet Cochran – Armitage testu. Tentokrát budeme volit v sekci **Table Statistics**, podsekci **Trend Test** a zde zaškrtneme volbu **Cochran-Armitage test** (obr. 21).



Obr. 21

Ve výsledcích si pak můžeme ověřit, že získáme stejnou hodnotu p-value pro SNP **`esr1_rs1042717`** jako v případě použití softwaru R (obr. 22)

Statistics for Table of Trait by esr1_rs1042717

Cochran-Armitage Trend Test	
Statistic (Z)	2.1205
One-sided Pr > Z	0.0170
Two-sided Pr > |Z|	0.0340

Effective Sample Size = 1000
Frequency Missing = 398

Obr. 22

3 Vícevýběrové testy kvantitativního znaku

3.1 Dvouvýběrový t-test

Dvouvýběrový t- test se užívá k testování nulové hypotézy o rovnosti středních hodnot ve dvou populacích.

$$H_0: \mu_1 = \mu_2$$

Můžeme např. definovat, že μ_1 je střední hodnotou populace jedinců s genotypem AA a μ_2 střední hodnotou populace jedinců s genotypy Aa a aa.

Důležitým předpokladem pro použití tohoto testu je, že výběry jsou nezávislé a mají normální rozdělení se stejným rozptylem (předpoklad stejného rozptylu se otestuje pomocí F-testu).

Testovací statistika pro t-test vypadá takto

$$T = \frac{\overline{X_n} - \overline{Y_m} - (\mu_1 - \mu_2)}{\sqrt{(n-1)S_n^2 + (m-1)S_m^2}} \sqrt{\frac{nm(n+m-2)}{n+m}} \sim t_{n+m-2},$$

kde $\overline{X_n}$, $\overline{Y_m}$ jsou výběrové průměry obou populací (populace s AA a populace s Aa,aa), S_n^2, S_m^2 jsou výběrové rozptyly a n, m jsou rozsahy výběrů.

Pro připomenutí výběrový průměr spočítáme jako

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

a výběrový rozptyl $S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$.

Za platnosti nulové hypotézy má testovací statistika Studentovo t-rozdělení o n+m–2 stupních volnosti. Je-li $|T| \geq t_{n+m-2}(1 - \frac{\alpha}{2})$, nulovou hypotézu H_O: $\mu_1 = \mu_2$ zamítneme ve prospěch H_A: $\eta_1 \neq \eta_2$.

3.1.1 Dvouvýběrový t-test v systému R

Chceme zjistit, zda alespoň jedna varianta alely pro nějaký SNP v resistin genu je asociována s proměnnou **NDRM.CH** (%-ní změnou síly nedominantního deltového svalu před a po cvičení).
Jako první, vytvoříme vektor názvů SNP v resistin genu a odpovídající matici genotypů.

```
> NamesResistinSnps = names(fms)[substr(names(fms),1,8)=="resistin"]
> fmsResistin = fms[,is.element(names(fms),NamesResistinSnps)]
```

Následně vytvoříme novou funkci, která konvertuje vektor genotypů na binární prvky (hodnoty 0 a 1) a vygeneruje p-hodnoty t-testu rovnosti středních hodnot znaků výsledných 2 skupin. K tomu bude potřeba nainstalovat a načíst balíček **genetics**. Definujeme také binární proměnnou genotypu **GenoBin** jako indikátor nejméně jedné varianty alely v odpovídajícím lokusu.

```
> library(genetics)
> TtestPval <- function(Geno){
+ alleleMajor <- allele.names(genotype(Geno, sep="", reorder="freq"))[1]
+ GenoWt <- paste(alleleMajor, alleleMajor, sep="")
+ GenoBin <- as.numeric(Geno!=GenoWt)[!is.na(Geno)]
+ Trait <- NDRM.CH[!is.na(Geno)]
+ return(t.test(Trait[GenoBin==1],Trait[GenoBin==0])$p.value)
+ }
```

Funkce **genotype()** mění objekt na typ objektu **genotype**, se kterým se v balíčku pracuje. Funkce **allele.names()** vyextrahuje z objektu typu **genotype** názvy alel. Volba **reorder="freq"** ve funkci **genotype()** zajistí změnu v pořadí alel dle četnosti výskytu alely. Do proměnné **alleleMajor** se tedy uloží hodnota nejčetnější alely.

Funkce **paste()** spojuje řetězce zadané jako její parametry. Volba **sep=""** zajistí, že mezi jednotlivými složkami řetězce se nebude vkládat žádný další znak. Funkce **is.na()** patří mezi informační funkce a vrací logickou hodnotu **TRUE** nebo **FALSE** podle toho, zda je hodnota jejího argumentu rovna konstantě **NA**, tj. chybějící hodnotě. Logický operátor „**!**“ má význam negace, proto také zápis „**!=**“ znamená nerovnost.

Nyní můžeme funkci **TtestPval** aplikovat na každý sloupec genotypové matice **fmsResistin**.

```
> apply(fmsResistin,2,TtestPval)
 resistin_c30t resistin_c398t resistin_g540a resistin_c980g resistin_c180g
    0.04401614     0.08098567     0.11578470     0.27828906     0.03969448
resistin_a537c
    0.06573061
```

Stanovíme-li opět hladinu významnosti na 0,05, pak dvouvýběrový t-test poukazuje na skutečnost, že první a pátý SNP v genu resistin mohou být asociovány s **NDRM.CH**.

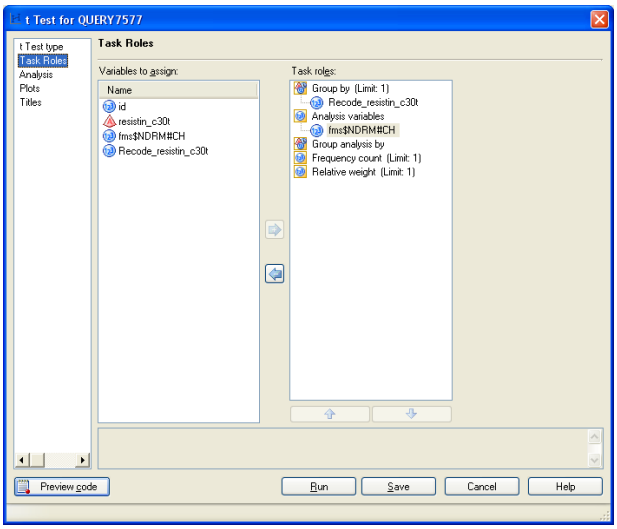
3.1.2 Dvouvýběrový t-test v systému SAS

Předpokládejme, že máme data připravená ve tvaru binární proměnné SNP (1=homozygot v nejčetnější alele, 0=jinak) – klasifikační proměnná, a zkoumané proměnné. Pro SNP **resistin_c30t** by mohla datová množina vypadat jako na obr. 23 (CC je homozygot, kde C je nejčetnější alela, **Recode_resistin_c30t** odpovídající binární klasifikační proměnná a **fms$NDRM#CH** analyzovaná proměnná).

	id	resistin_c30t	fms$NDRM#CH	Recode_resistin_c30t
1	1	CC	40	1
2	2	CC	25	1
3	3	CC	40	1
4	4	CC	125	1
5	5	CT	40	0
6	6	CC	75	1
7	7	CC	100	1
8	8	CC	.	1
9	9	CC	57.1	1
10	10	CC	33.3	1
11	11	CC	.	1
12	12	CC	20	1
13	13	CC	25	1
14	14	CC	100	1
15	16	CC	28.6	1
16	17	CC	.	1
17	18	CC	.	1
18	19	CC	7.1	1
19	20	CC	75	1
20	21	CC	12.5	1
21	22	CC	60	1
22	23	CC	40	1

Obr. 23

Analýzu prostřednictvím Studentova t-testu provedeme pomocí úlohy **t Test** v nabídce **Analyze**, podnabídce **ANOVA**. Nastavíme role dle obr. 24. Jinak nemusíme nic nastavovat.



Obr. 24

Ve výsledcích uvidíme jak F-test pro shodu rozptylů, tak jednotlivé varianty t-testu (pro shodné rozptyly i aproximaci pro různé rozptyly). Poněvadž F-test vyšel významně (p-value = 0.0154), použijeme jako výsledek t-testu řádek označený jako Satterthwaite (ve sloupci Variances je uvedeno Unequal) a zjistíme, že p-hodnota pro oboustranný test je rovna číslu 0.044, což odpovídá výsledkům získaným pomocí softwaru R.

The TTEST Procedure

Statistics

Variable	Recode_resistin_c30t	N	Lower CL Mean	Mean	Upper CL Mean	Lower CL Std Dev	Std Dev	Upper CL Std Dev	Std Err	Minimum	Maximum
fms$NDRM#CH	0	17	32.575	42.641	52.707	14.581	19.577	29.795	4.7482	10	100
fms$NDRM#CH	1	585	50.605	53.315	56.025	31.566	33.375	35.406	1.3799	0	250
fms$NDRM#CH	Diff (1-2)		-26.66	-10.67	5.3113	31.311	33.082	35.066	8.1392		

T-Tests

Variable	Method	Variances	DF	t Value	Pr > |t|
fms$NDRM#CH	Pooled	Equal	600	-1.31	0.1902
fms$NDRM#CH	Satterthwaite	Unequal	18.8	-2.16	0.0440

Equality of Variances

Variable	Method	Num DF	Den DF	F Value	Pr > F
fms$NDRM#CH	Folded F	584	16	2.91	0.0154

Obr. 25

3.2 Wilcoxonův dvouvýběrový test

Wilcoxonův dvouvýběrový test, známý též jako Mannův-Whitneyův test je neparametrickou analogií dvouvýběrového t-testu. Jeho použití je vhodné, pokud máme malé rozsahy výběrů (n_1,n_2) nebo není zajištěna normalita výběrů.
Testujeme hypotézu, že rozdělení obou populací je shodné.

Postupujeme tak, že seřadíme hodnoty obou populací od nejnižších po nejvyšší a určíme pořadí. Potom provedeme součty pořadí v jednotlivých populacích. Součet pořadí v populaci 1 označíme jako S_1 a součet pořadí v druhé populaci S_2.

Nyní můžeme spočítat hodnoty testovací statistiky U_1 a U_2

$$U_1 = n_1 n_2 - \frac{n_1(n_1+1)}{2} - S_1 \qquad U_2 = n_1 n_2 - \frac{n_2(n_2+1)}{2} - S_2$$

Pro U_1 a U_2 platí vztah, že $U_1 + U_2 = n_1 n_2$, lze tedy vypočítat jen jednu ze statistik a druhou dopočítat z tohoto vztahu.

Menší z hodnot statistik U_1 a U_2 porovnáme s kritickou hodnotou pro dvouvýběrový Wilcoxonův test. Pokud $\min(U_1,U_2)$ je menší nebo rovno kritické hodnotě, zamítneme nulovou hypotézu, tedy rozdělení v populacích není stejné.

Speciálně v případě, kdy jsou rozsahy výběrů n_1 a n_2 velké (alespoň 20) vypočteme veličinu Z, která má normované normální rozdělení N(0,1)

$$Z = \frac{U_1 - \frac{1}{2}n_1 n_2}{\frac{1}{12}n_1 n_2 (n_1 + n_2 + 1)}.$$

Vypočtenou hodnotu z porovnáme s kritickou hodnotou, $(1-\frac{\alpha}{2})$-kvantilem normálního rozdělení $z(1-\frac{\alpha}{2})$Pokud je absolutní hodnota z rovna nebo větší jak $z(1-\frac{\alpha}{2})$, zamítneme hypotézu H_0 na hladině významnosti α.

3.2.1 Wilcoxonův dvouvýběrový test v R

Provádí se podobně jako dvouvýběrový t-test, s tím rozdílem, že funkci **`t.test()`** nahradíme funkcí **`wilcox.test()`.**

```
> attach(fms)
> NamesResistinSnps = names(fms)[substr(names(fms),1,8)=="resistin"]
> fmsResistin = fms[,is.element(names(fms),NamesResistinSnps)]
> library(genetics)

> WilcoxTPval = function(Geno){
+ alleleMajor = allele.names(genotype(Geno, sep="",
+ reorder="freq"))[1]
+ GenoWt = paste(alleleMajor, alleleMajor, sep="")
+ GenoBin = as.numeric(Geno!=GenoWt)[!is.na(Geno)]
+ Trait = NDRM.CH[!is.na(Geno)]
+ return(wilcox.test(Trait[GenoBin==1], Trait[GenoBin==0])$p.value)
+ }
```

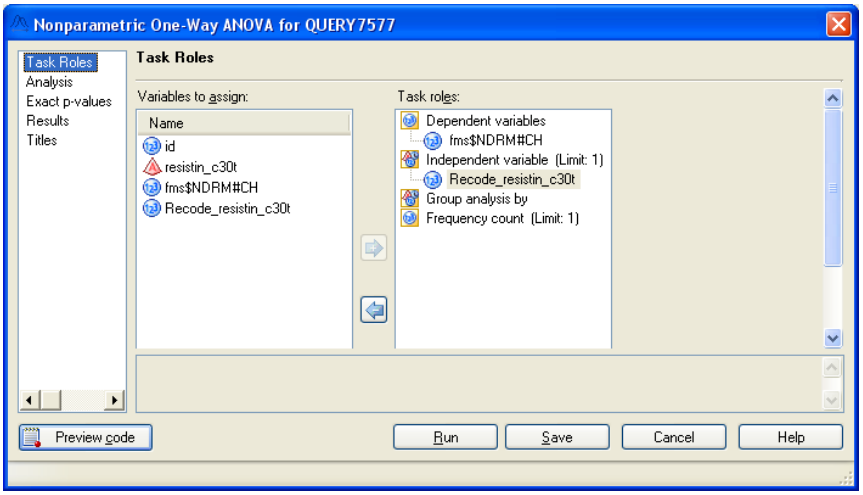
Aplikujeme funkci **`WilcoxTPval`**.

```
> apply(fmsResistin,2,WilcoxTPval)

resistin_c30t resistin_c398t resistin_g540a resistin_c980g resistin_c180g
    0.25905754     0.08348861     0.05177213     0.69136163     0.02981086
resistin_a537c
    0.21398246
```

Wilcoxonův test na rozdíl od dvouvýběrového t-testu vypovídá o tom, že pouze pátý SNP v resistin genu může být asociován s **`NDRM.CH`** (zde je p-hodnota menší než α).

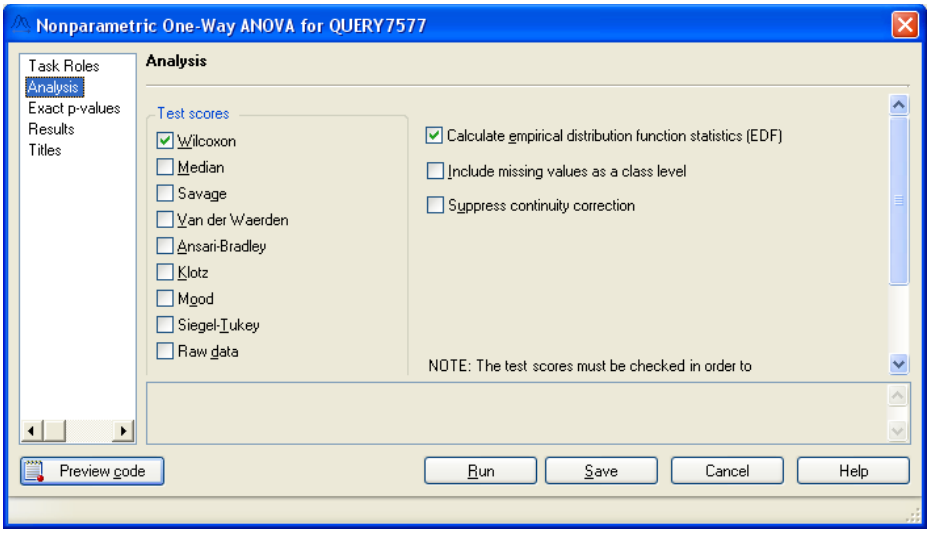
3.2.2 Wilcoxonův dvouvýběrový test v systému SAS

Pro stejnou datovou množinu jakou jsme použili u t-testu, provedeme nyní analýzu prostřednictvím Wilcoxonova neparametrického testu. K tomuto účelu využijeme úlohy **Nonparametric One-Way ANOVA**, která se skrývá ve stejné nabídce jako úloha **t Test**, tj. **Analyze**, **ANOVA.** Nastavení rolí je uvedeno na obr. 26.



Obr. 26

V sekci **Analysis** ponecháme zatrženou pouze volbu **Wilcoxon** (obr. 27).



Obr. 27

Ve výsledcích poté zjistíme p-hodnotu oboustranného dvouvýběrového Wilcoxonova testu, která odpovídá hodnotě získané prostřednictvím softwaru R, tj. 0.2591 (viz obr. 28).

Wilcoxon Two-Sample Test	
Statistic	4328.0000
Normal Approximation	
Z	-1.1286
One-Sided Pr < Z	0.1295
Two-Sided Pr > |Z|	0.2591
t Approximation	
One-Sided Pr < Z	0.1298
Two-Sided Pr > |Z|	0.2595
Z includes a continuity correction of 0.5.	

Obr. 28

3.3 Analýza rozptylu (ANOVA)

Analýza rozptylu nám umožňuje porovnávat více nezávislých výběrů (můžeme porovnávat populace s genotypy AA, aa, Aa) a testovat hypotézu o rovnosti středních hodnot v *k* výběrech oproti alternativní hypotéze, kdy alespoň dva populační průměry se od sebe liší.

$$H_0: \mu_1 = ... = \mu_k \qquad (5)$$

Důležitým předpokladem pro užití analýzy rozptylu je, že výběry mají normální rozdělení se stejnou směrodatnou odchylkou.

Matematický model se zapisuje takto

$$Yij = \mu + \alpha_i + \varepsilon_{ij}$$

j-té pozorování i-té skupiny

Společná střední hodnota

Efekt faktoru A v i-tém výběru

Náhodné chyby

Variabilita uvnitř skupin určuje, jak se hodnoty v jednotlivých skupinách liší od průměru skupiny. Vypočteme ji pomocí reziduálního součtu čtverců S_E.
Nejprve určíme skupinový průměr pro všechny skupiny dle vztahu

$$\overline{Y_{i.}} = \frac{1}{n_i}\sum_{j=1}^{n} Y_{ij} ,$$

kde n_i je počet pozorování v i-té skupině (výběru). Potom můžeme spočítat S_E jako sumu druhých mocnin rozdílů hodnot a příslušného skupinového průměru

$$S_E = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y}_{i.})^2.$$

Variabilitu jednotlivých pozorování kolem celkového průměru (průměru všech pozorování) charakterizuje celkový součet čtverců S_T.
Nejdříve musíme určit celkový průměr.

$$\overline{Y_{..}} = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i} Y_{ij} = \frac{1}{n}\sum_{i=1}^{k} n_i \overline{Y}_{i.} ,$$

kde n je celkový počet pozorování (ve všech skupinách).

S_T určíme tak, že od každé hodnoty pozorování odečteme celkový průměr a výsledné rozdíly umocníme na druhou

$$S_T = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y_{..}})^2.$$

Variabilita mezi skupinami ukazuje, jak se liší skupinové průměry od celkového průměru. Zjistíme ji výpočtem skupinového součtu čtverců S_A

$$S_A = S_T - S_E.$$

Porovnání variabilit uvnitř a mezi skupinami provedeme pomocí F-testu. Vypočtené hodnoty dosadíme do testovací statistiky F_A, která má za platnosti nulové hypotézy F rozdělení o k - 1, n - k stupních volnosti, kde k je počet skupin a n je celkový počet pozorování

$$F_A = \frac{S_A}{S_E}\frac{n-k}{k-1}.$$

Pokud hodnota testovací statistiky $F_A \geq F_{k-1,n-k}(1-\alpha)$, zamítneme hypotézu (5) na hladině významnosti α.

Výsledky výpočtů se pro přehlednost shrnují do tabulky typické pro analýzu rozptylu.

Zdroj variability	Součet čverců	Stupně volnosti	F_A	p-hodnota
Skupiny	S_A	k-1	$\frac{S_A}{S_E}\frac{n-k}{k-1}$	$P(F_{k-1,n-k} \geq F_A)$
Reziduální	S_E	n-k		
Celkový	S_T	n-1		

3.3.1 Analýza rozptylu v R

Chceme odhalit asociaci mezi **resistin_c180g** SNP a procentní změnou síly nedominantního svalu před a po cvičení (**NDRM.CH**).

Začneme načtením genotypových dat – proměnné **resistin_c180g** SNP a definováním znaku. Funkce **as.factor()** změní typ proměnné na faktor.

```
> attach(fms)
> Geno = as.factor(resistin_c180g)
> Trait = NDRM.CH
```

K provedení analýzy rozptylu užijeme funkce **lm()** . Nejprve nadefinujeme, že chceme vyloučit jedince s chybějícími hodnotami znaku (v souboru dat značeno **NA**). Toho dosáhneme pomocí podmínky **na.action==na.exclude**. Funkcí **summary()** vyvoláme výstup, jehož součástí je tabulka v obvyklém formátu pro ANOVA test.

Alternativně lze provést pomocí funkce **aov()**, s tím rozdílem, že výstup vyvoláme příkazem **print()**.

```
> AnovaMod = lm(Trait ~ Geno,na.action=na.exclude)
> summary(AnovaMod)
```

```
Call:
lm(formula = Trait ~ Geno, na.action = na.exclude)

Residuals:
    Min      1Q  Median      3Q     Max
-56.054 -22.754  -6.054  15.346 193.946

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   56.054      2.004  27.973   <2e-16 ***
GenoCG        -5.918      2.864  -2.067   0.0392 *
GenoGG        -4.553      4.356  -1.045   0.2964
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 33.05 on 603 degrees of freedom
  (791 observations deleted due to missingness)
Multiple R-squared: 0.007296,   Adjusted R-squared: 0.004003
F-statistic: 2.216 on 2 and 603 DF,  p-value: 0.1100
```

Protože p-hodnota je větší než α, nemůžeme zamítnout hypotézu o rovnosti středních hodnot (5).

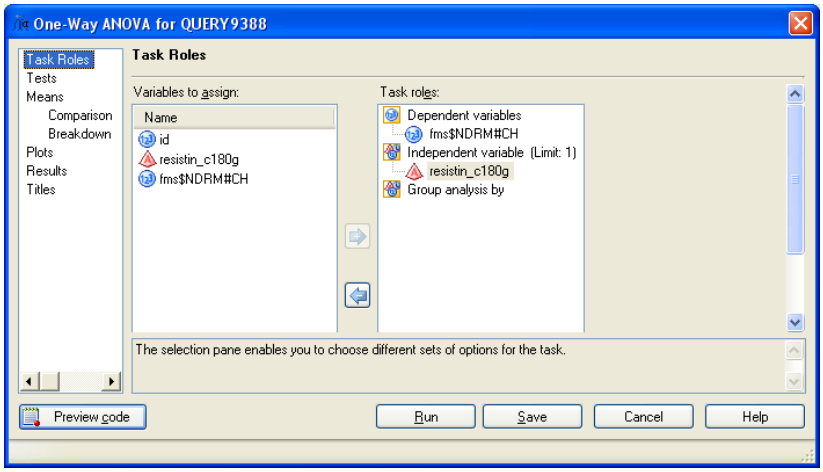
3.3.2 ANOVA v systému SAS

Předpokládejme, že máme k dispozici datovou množinu odpovídající klasifikační proměnné SNP **`resistin_c180g`** a proměnnou analyzovanou, v našem případě **`NDRM.CH`** (viz obr. 29).

	id	resistin_c180g	fms$NDRM#CH
1	1	CC	40
2	2	GG	25
3	3	CC	40
4	4	CG	125
5	5	CC	40
6	6	CC	75
7	7	CG	100
8	8	CG	.
9	9	CC	57.1
10	10	CC	33.3
11	11	CG	.
12	12	CG	20

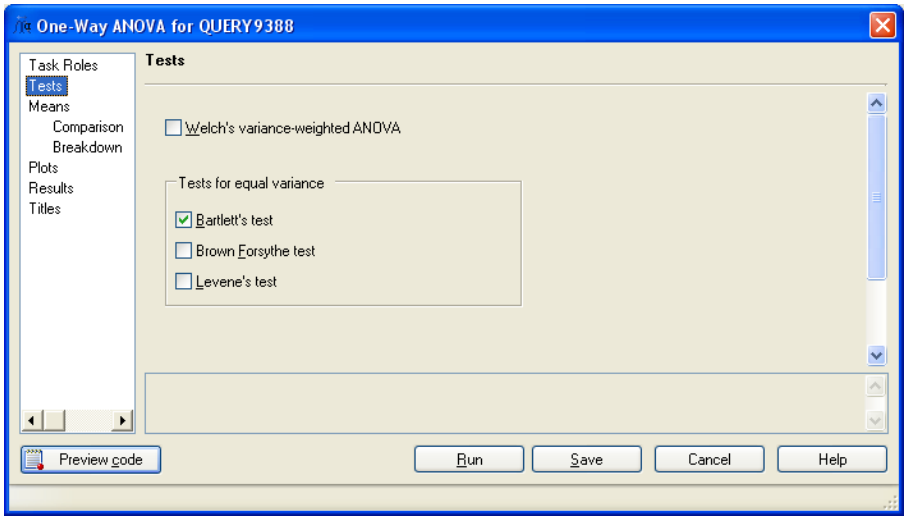
Obr. 29

Pro jednocestnou analýzu rozptylu je v SAS EG k dispozici přímo úloha **One-Way ANOVA** v nabídce **Analyze**, podnabídce **ANOVA**. Přiřazení rolí provedeme dle obr. 30.



Obr. 30

V sekci **Tests** můžeme zaškrtnout Bartlettův test k ověření hypotézy, že výběry pochází z normálně rozdělených souborů se stejným rozptylem (ověření homoskedasticity), viz obr. 31.



Obr. 31

Jinak nemusíme nic nastavovat. Ve výsledcích vidíme, že Bartlettův test nezamítl hypotézu shody rozptylů v jednotlivých skupinách a můžeme také porovnat p-hodnotu získanou v SAS EG s tou, kterou jsme získali prostřednictvím softwaru R (obr. 32).

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	4840.1316	2420.0658	2.22	0.1100
Error	603	658597.3915	1092.2013		
Corrected Total	605	663437.5231			

Bartlett's Test for Homogeneity of fms$NDRM#CH Variance			
Source	DF	Chi-Square	Pr > ChiSq
resistin_c180g	2	3.1008	0.2122

Obr. 32

3.4 Kruskalův – Wallisův test

Kruskalův-Wallisův test je vlastně neparametrickou obdobou jednofaktorové analýzy rozptylu. Jde o zobecnění Wilcoxonova dvouvýběrového testu pro *k* výběrů. Jeho použití je vhodnější u výběrů s malým rozsahem. U této metody nemusí být splněn předpoklad normality výběrů, budeme proto předpokládat, že každý z *k* výběrů pochází z rozdělení se spojitou distribuční funkcí a otestujeme, zda výběry pochází ze stejného rozdělení.

Postupujeme tak, že seřadíme všech n prvků od nejmenšího po největší a přiřadíme prvkům pořadí. Potom spočteme součet pořadí i-tého výběru T_i , kde i = 1,…k a následně určíme hodnotu veličiny Q

$$Q = \frac{12}{n(n+1)} \sum_{i=1}^{k} \frac{T_i^{\,2}}{n_i} - 3(n+1) ,$$

kde n_i je rozsah i-tého výběru, i = 1,…k a n je $\sum_{i=1}^{k} n_i$.

Za platnosti hypotézy, že výběry pochází z téhož rozdělení má veličina Q asymptotické χ^2 rozdělení s počtem stupňů volnosti k – 1. Je-li $|q| \geq \chi^2_{k-1}(1-\alpha)$, zamítáme hypotézu na hladině významnosti α.

3.4.1 Kruskalův – Wallisův test v R

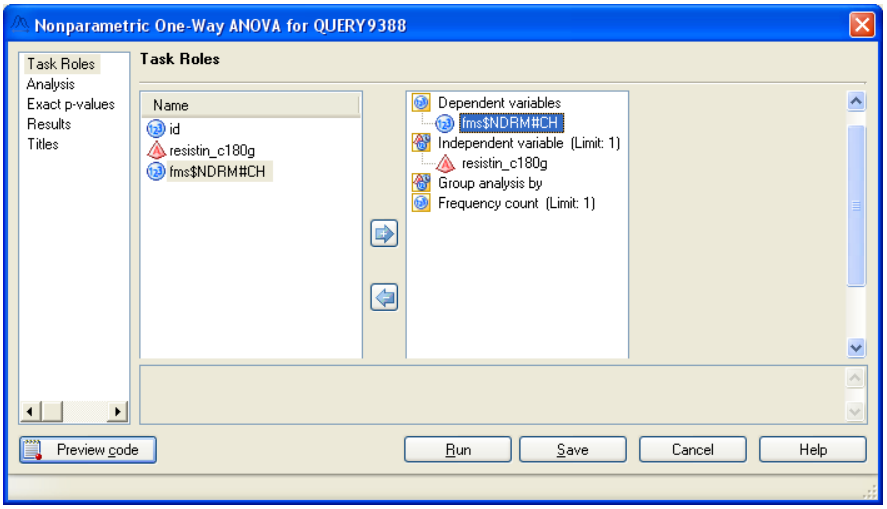
Použijeme, pokud máme výběry menšího rozsahu, ve kterých nemusí být splněn předpoklad normality.
Opět musíme specifikovat, že chceme, aby byli vyloučeni jedinci s chybějícími hodnotami znaku.

```
> kruskal.test(Trait, Geno, na.action=na.exclude)

        Kruskal-Wallis rank sum test

data:  Trait and Geno
Kruskal-Wallis chi-squared = 4.9268, df = 2, p-value = 0.08515
```

Můžeme pozorovat, že jsme dosáhli stejného výsledku – není prokázána asociace mezi **resistin_c180g** a **NDRM.CH**, jelikož p-hodnota je větší jak stanovená hladina významnosti α.

3.4.2 Kruskalův-Wallisův test v systému SAS

Pro stejnou datovou množinu jako u testu ANOVA můžeme provést neparametrickou obdobu tohoto testu, tj. Kruskalův – Wallisův test. V SAS EG k tomuto účelu slouží úloha **Nonparametric One-Way ANOVA**, kterou jsme použili již pro dvouvýběrový Wilcoxonův test. Nastavení rolí pro případ SNP **resistin_c180g** a analyzovanou proměnnou **NDRM.CH** je na obr. 33.



Obr. 33

V části **Analysis** necháme zaškrtnutou pouze volbu **Wilcoxon**. Ve výsledcích potom vidíme p-hodnotu Kruskalova – Wallisova testu, podobnou jako při výpočtu v softwaru R (obr. 34).

Wilcoxon Scores (Rank Sums) for Variable fms$NDRM#CH Classified by Variable resistin_c180g

resistin_c180g	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
CC	272	87205.00	82552.00	2141.44898	320.606618
GG	73	21737.50	22155.50	1401.44185	297.773973
CG	261	74978.50	79213.50	2131.96391	287.273946

Average scores were used for ties.

Kruskal-Wallis Test	
Chi-Square	4.9268
DF	2
Pr > Chi-Square	0.0851

Obr. 34

Závěr

Myslím, že v mé práci se úspěšně podařilo vysvětlit čtenáři princip jednotlivých vybraných statistických metod v populačních asociačních studiích a jejich zpracování v softwarech SAS a R.

Psaní práce bylo pro mě velice zajímavé, jelikož jsem nahlédla při aplikacích metod, které jsem se naučila během mého bakalářského studia, také do jiné oblasti, oblasti genetiky. Dalším přínosem mé práce bylo naučit se lépe pracovat v prostředí systémů SAS a R a s jejich pomocí vypočítat hodnoty testovacích statistik testů asociace jednonukleotidového polymorfismu.

Vzhledem k rozsáhlé povaze tématu se však nepodařilo obsáhnout všechny metody používané v této oblasti, což je velká škoda, jelikož populační asociační studie jsou velmi zajímavou částí genetiky. Čtenáře proto mohu odkázat na literaturu [1], kde se dozví vše, co se do této bakalářské práce nevešlo. Pro mě je to velkou výzvou, že bych se mohla podrobným vysvětlením zabývat v diplomové práci magisterského studia.

Seznam použité literatury:

[1] FOULKES, Andrea S. *Applied Statistical Genetics with R : For Population-based Association Studies*. New York (USA) : Springer Dordrecht Heidelberg London New York, 2009. 252 s. ISBN 978-0-387-89553-6, e-ISBN 978-0-387-89554-3. DOI 10.1007/978-0-387-89554-3.

[2] FIALA, Jaromír. *Biologie IV. : Genetika, pracovní sešit pro multimediální výuku biologie*. Vydání první. Boskovice : František Šalé - ALBERT, 2005. 62 s. ISBN 80-7326-063-8.

[3] SECKO, David. *The Science Creative Quarterly* [online]. 2003 [cit. 2010-04-27]. A MONK'S FLOURISHING GARDEN: THE BASICS OF MOLECULAR BIOLOGY EXPLAINED. Dostupné z WWW: <http://www.scq.ubc.ca/a-monks-flourishing-garden-the-basics-of-molecular-biology-explained/>.

[4] *Cytogenetická laboratoř Brno* [online]. 2008 [cit. 2010-04-27]. Zajímavosti. Dostupné z WWW: <http://www.cytogenetika.cz/zajimavosti.html>.

[5] HANČOVÁ, Hana ; VLKOVÁ, Marie. *BIOLOGIE V KOSTCE I*. 1.vydání. Havlíčkův Brod : FRAGMENT, 1997. 112 s. ISBN 80-7200-059-4.

[6] PANČÍK, Peter; MARCIŠOVÁ, Denisa. *BIOWEB* [online]. 2003-2009 [cit. 2010-04-27]. Génové mutácie. Dostupné z WWW: <http://www.bioweb.genezis.eu/index.php?cat=7&file=genovemut>.

[7] PANČÍK, Peter; MARCIŠOVÁ, Denisa. *BIOWEB* [online]. 2003-2009 [cit. 2010-04-27]. Základné genetické pojmy. Dostupné z WWW: <http://www.bioweb.genezis.eu/?cat=7&file=pojmy>.

[8] RELICHOVÁ, Jiřina. *Genetika populací*. 1. vydání. Brno : Masarykova univerzita/Nakladatelství pro Přírodovědeckou fakultu, Tisk Coprint, 2009. 187 s. ISBN 978-80-210-4795-2.

[9] ZVÁROVÁ, Jana; MAZURA, Ivan; SVATOŠ, Jan. *STOCHASTICKÁ GENETIKA*. 1. vydání. Praha : Univerzita Karlova v Praze - Nakladatelství Karolinum, 2001. 171 s. ISBN 80-246-0264-4.

[10] *Aktuální genetika : Multimediální učebnice lékařské biologie, genetiky a genomiky* [online].Ústav biologie a lékařské genetiky 1.LF UK a VFN, 2005-2006 [cit. 2010-04-27]. Stručný slovník / glosář genetických pojmů. Dostupné z WWW: <http://biol.lf1.cuni.cz/ucebnice/slovnik.htm>.

[11] *Wikiskripta* [online]. 2009, 5. 12. 2009 [cit. 2010-04-27]. Hardy-Weinbergova rovnováha. Dostupné z WWW: <http://www.wikiskripta.eu/index.php/Hardy-Weinbergova_rovnov%C3%A1ha>.

[12] Radka Storchová - Základy genetiky [online]. 2009-2010 [cit. 2010-04-28]. Evoluční genetika. Dostupné z WWW: <http://web.natur.cuni.cz/~radkas/prezentace/EvolucniGenetika.pdf>.

[13] ZVÁROVÁ, Jana ; MALÝ, Marek. *STATISTICKÉ METODY V EPIDEMIOLOGII - svazek 2*. 1.vydání. Praha : Univerzita Karlova v Praze - Nakladatelství Karolinum, 2003. 505 s. ISBN 80-246-0764-6 (svazek 2), ISBN 80-246-0765-4 (soubor).

[14] ZVÁROVÁ, Jana ; MALÝ, Marek. *STATISTICKÉ METODY V EPIDEMIOLOGII - svazek 1*. 1.vydání. Praha : Univerzita Karlova v Praze - Nakladatelství Karolinum, 2003. 236 s. ISBN 80-246-0763-8 (svazek 1), ISBN 80-246-0765-4 (soubor).

[15] *Wikipedie, otevřená encyklopedie* [online]. 2010, 24. 4. 2010 v 21:41 [cit. 2010-04-27]. Genom. Dostupné z WWW: <http://cs.wikipedia.org/wiki/Genom>.

[16] *Velký lékařský slovník on-line* [online]. 2008 [cit. 2010-04-27]. Genom. Dostupné z WWW: <http://lekarske.slovniky.cz/pojem/genom>.

[17] ZVÁROVÁ, Jana. *ZÁKLADY STATISTIKY pro biomedicínské obory*. 1.vydání. Praha : Univerzita Karlova v Praze - Nakladatelství Karolinum, 1998. 218 s. ISBN 80-7184-786-0.

[18] *Wikipedie, otevřená encyklopedie* [online]. 2010, 25. 4. 2010 v 06:30 [cit. 2010-04-27]. Testování statistických hypotéz. Dostupné z WWW: <http://cs.wikipedia.org/wiki/Testov%C3%A1n%C3%AD_statistick%C3%BDch_hypot%C3%A9z>.

[19] ANDĚL, Jiří . *ZÁKLADY MATEMATICKÉ STATISTIKY*. Vydání první. Praha : MATFYZPRESS, vydavatelství Matematicko-fyzikální fakulty Univerzity Karlovy v Praze, 2005. 358 s. ISBN 80-86732-40-1.

[20] KUNDEROVÁ, Pavla. *ZÁKLADY PRAVDĚPODOBNOSTI A MATEMATICKÉ STATISTIKY*. 1. vydání. Olomouc : Univerzita Palackého v Olomouci, 2004. 184 s. ISBN 80-244-0813-9.

[21] ANDĚL, J. MATEMATICKÁ STATISTIKA. Praha : SNTL/Alfa, 1978.

[22] *Pravděpodobnost a statistika HYPERTEXTOVĚ* [online]. 2004 [cit. 2010-04-29]. P-hodnota. Dostupné z WWW: <http://home.zcu.cz/~friesl/hpsb/phodn.html>.