記述統計量，仮説検定の基本

- http://minato.sip21c.org/grad/infop-text2012.pdf（pp.8-19を印刷して配布します）
- 平均値(Mean)と中央値(Median)
- 標準偏差(Standard Deviation)と標準誤差(Standard Error)
- 仮説検定(Hypothesis testing)
  - 平均値の比較
  - 比率（割合）の比較

標準偏差と標準誤差

- 標準偏差(SD)は，データそのもののばらつき
  - SD = √Σ (xi – x)^2/(n-1)
- 標準誤差(SE)は推定値のばらつき
  - SEM (standard error of mean) = SD/√n
  - 研究を何度も繰り返して行えば，推定値のばらつきは，ある確率でSEの幅に入る
  - （例）回帰分析では，傾き（回帰係数）の標準誤差を計算することができる

仮説検定の仕組み

- 2群間で差があるとか，2つの変数の間に相関があると言いたい場合，どの程度の差や相関なのかを特定できないので，「差が無い」「相関が無い」という帰無仮説を立てる
- 帰無仮説（棄却されるべき仮説）のもとで，期待値から現在のデータ以上に外れたデータが偶然得られる確率を計算する
- もし，その確率が有意水準（予め決めておかねばならない）より小さかったら，我々は「帰無仮説が間違っていた」と判定することができる。そうでない場合は，帰無仮説についての判定を保留する。

平均値の比較

- 2群間で平均値を比較するt検定は，ざっくり言えば分散分析の特殊な場合（群数が2の場合）
  - 検定統計量としてF値でなくt値を使うが数学的に同等
  - 2群間で等分散を仮定せずに分析するには，Welchの方法で自由度を調整する
- 分散分析表(ANOVA table)は，分散を級内分散と級間分散に分解する
  - もし，級間分散が級内分散に比べて遙かに大きければ（つまり分散比であるF値が1より遙かに大きければ），「級分け変数が値に有意に影響している」と判断できる

割合（比率）の比較 (1)

- A, B, Cという3つの処理を実施したときの脱落が
  A: 0/15, B: 2/13, C: 1/14であったとき，
  「3つの処理間で脱落率に差が無い」という帰無仮説の下で期待される脱落率は，
  (0+2+1)/(15+13+14) =3/42.
  これが正しければ，期待されるクロス表は

	A	B	C
脱落	15*(3/42)	13*(3/42)	14*(3/42)
完了	15*(39/42)	13*(39/42)	14*(39/42)

- すべてのセルについて，観測度数と期待度数の差の2乗を期待度数で割った値を合計したものがカイ二乗値X2
  X2 = (0-15*(3/42))^2/(15*(3/42)) + … = 2.485…
- X2は自由度2（3処置，2種類の結果だから，それぞれから1を引いて積をとったものが自由度）のカイ二乗分布に従う。1-pchisq(X2, 2)を計算すると（注：pchisq(X2, 2)は自由度2のカイ二乗分布のX2までの確率密度の積分値），0.2886
  - Rではprop.test(c(0,2,1), c(15,13,14), correct=FALSE)で
    X-squared = 2.4852, df = 2, p-value = 0.2886
    と結果が得られる。

割合（比率）の比較 (2)

- この例では脱落数が少ないので，カイ二乗検定では近似が悪い→どうする？
- 「処理間で脱落率に差が無い」という帰無仮説は，「脱落するかどうかは処理の種類と無関係（＝独立）」という帰無仮説と同値なので，以下のクロス表を想定し，フィッシャーの正確な検定(Fisher's exact test)を実行

	A	B	C	Total
Dropout	0	2	1	3
Complete	15	11	13	39
Total	15	13	14	42

- 「脱落するかどうか」と「処理の種類」が偶然この組合せになっている確率を計算するには，この表と同じ周辺度数をもつクロス表（例えば下記）をすべて考え，それぞれが偶然得られる確率を計算。下表の確率は${}_{15}C_1 * {}_{13}C_1 * {}_{14}C_1 / {}_{42}C_3$で，約0.238となる。実際の脱落データの表は，${}_{15}C_0 * {}_{13}C_2 * {}_{14}C_1 / {}_{42}C_3$で，約0.095となる。

	A	B	C	Total
Dropout	1	1	1	3
Complete	14	12	13	39
Total	15	13	14	42

- 実際の表より偶然得られる確率が小さな表の確率を合計→フィッシャーの正確な確率が得られる。0.095+0.040+0.025+0.032=0.192
  Rではfisher.test(matrix(c(0,15,2,11,1,13),2,3)で0.1914 （違いは丸め誤差）