実験データの取得・大規模試験データの収集

データ取得・収集の基礎

集めたデータがデータとして使えるようにするには

以下の２つの項目を記録することが必要

- データレコード(サンプル7の検査Xの値は32)そのもの
- データレコードの集合体に関する情報(メタ情報)

データ取得・収集の記録

データ取得・収集について、以下の２点について、記録することは必須であり、学術・研究の基礎(研究記録)

- いつだれが行ったかがわかるようにする
- 再現性を担保する

この２点は、データ収集に限らず、データ解析の実施等、すべての活動に共通する

臨床記録(カルテ記録)も同様

実習

実験(等)はすべて自分で行ったものとして記録を作成すること

実験１

さいころ製造工場では、製造したさいころの精度管理のために抜き打ち検査をしています。

2019年7月3日に、12台ある製造機のうちの第8番製造機にて午前８時半から午前９時の間に製造されたさいころをランダムに１個抜き出し、以下の実験をしました。

なお、この日のこの時刻に第8番製造機では、材料Aについては、X社製の製品ロット7X32が用いられ、同機の運転と、製造さいころの視認選別(目で見て不良品と思われるさいころは、製造ラインにて選別)とを担当したのは、社員Bでした。 また、抜き打ち検査にてさいころを振り、その出た目を記録したのは自分であるものとする。

```
dprob <- rep(1/6,6) + rnorm(6,0,0.03)
dprob <- dprob/sum(dprob)
n.trials <- 20
res <- rep(NA,n.trials)
for(i in 1:n.trials){
  num <- sample(1:6,1,prob=dprob)
  print(num)
  res[i] <- num
}
```

```
## [1] 2
## [1] 4
## [1] 2
## [1] 2
## [1] 3
## [1] 1
## [1] 5
## [1] 4
## [1] 6
## [1] 2
## [1] 3
## [1] 2
## [1] 5
## [1] 1
## [1] 5
## [1] 2
## [1] 6
## [1] 1
## [1] 4
## [1] 5
```

講義内課題 1

上記の実験記録を手書きで作りなさい。

作成時間は５分。

周囲の受講者と相談してもよいが、記録を書くのは自分自身とすること。

また、コンピュータを立ち上げ、アプリケーションRを起動すれば、同様の実験が可能であるので、自身で実験した結果を用いて実験記録を書いてもよい。

相互チェックと、修正の仕方

周囲の受講者２名と記録を交換し、自分の記録と他人の記録の異同を確認し、改善したい点があれば、修正しなさい。

ただし、データの記録は消しゴムで消したり、ホワイトで修正したりすることは許されないものなので、以下に注意すること

- 抹消したい部分は二重線で、何を消したかがわかるように消し、加筆についても、加筆であることがわかるように行うこと
- また、修正の際には、誰がいつ修正したかも重要な情報であるので、署名(イニシャル等で修正者が特定できるのであれば、イニシャルでもよい)。また、加除訂正の理由も必要に応じて記録すること
- なお、これは、記録とその修正の実習であって、記録の記載を評価することが目的ではないので、上記注意を理解し、実践することに意味があることに注意して、修正すること

講義内課題 2

ある病気に対する２種類の治療法に関する生存解析 (http://jspt.japanpt.or.jp/ebpt_glossary/survival-analysis.html)を行っているとします。

患者さんがエントリーするたびに、個人を識別できる情報(名前、カルテ番号、生年月日、住所など)は分からないようにして(匿名化)、患者ID(患者識別子)とともに、以下の情報が、データ管理室に届くことになっている

- 患者ID

- 割付治療タイプ {A,B}
- 登録日
- 治療開始日
- (疾病情報：癌のステージ等。今回は省きます)
- (共変量情報：性別・年齢等。今回は省ぎます)

また、患者さんが死亡したことが確認された場合は、その時点で、死亡日の情報が届きます

また、定期的に患者さんは追跡調査され、その時点で生存しているか否かの情報が届きます。 このような患者さんを追跡調査する場合には、何かしらの理由で、追跡不能となることがあります(引っ越し等)。スタディからの脱落と言います。

また、副作用等の発生のために、スタディから外れることもあります。これも脱落ですが、治療法と関連の深い脱落なので、上記のような単なる追跡不能とは区別することが適切です。

追跡はできているが、生存・死亡の情報が医療機関から届かないこともあり、別の意味での欠測値として扱うことが適切です。

上記の情報は、いずれも、いつ、だれが、データ管理室に届けたかも大切な情報です。

あなたは、データ管理室で働いており、届くデータの記録を取る仕事を担当しています。

配布資料のように届いたデータを整理してエクセルファイルにまとめなさい。

相互チェックと、修正の仕方

周囲の受講者２名と記録を見せあいし、自分の記録と他人の記録の異同を確認し、改善したい点があれば、修正しなさい。

ただし、ファイルで作成したデータの記録はファイルの中身を書き換えて行ってはいけません。ファイルを修正しても、その修正ログをきちんと残す仕組みが世の中にはすでにあり、それを使うことが、推奨されますが、必ずしも、皆さんがそのような仕組みを使いこなすようになるとは限らないので、この実習では、次のようにしてファイルを書き換えずに修正する方法を実習します。

ファイル名にバージョンを導入する。

hoge1.xlsx hoge2.xlsx …

このとき、修正の理由など必要な情報は、備考用のエクセルシートに書いてもよいですし、 データレコードと同じシートに備考を書き込むように作ることもできます。

また、エクセルなどの電子ファイルは、「修正不可」としておけば、作成日時がファイルの属性として記録されます。

作成日時の確認方法も覚えておきます。

- なお、これは、記録とその修正の実習であって、記録の記載を評価することが目的ではないので、上記注意を理解し、実践することに意味があることに注意して、修正すること

大規模データの取得

臨床系データでは少し遅れていますが、実験系データでは、大規模に取得したデータは原則、公共データベースに置いて、ダウンロードすることができるようにすることとなっています。

- 解析の再現性を他者が確認できるようにする(研究の透明性の確保)
- 公的研究資金を用いて得られたデータは公共財であるという考え方から、公開することが当然であると考えられる

GEO (Gene expression omnibus) (https://www.ncbi.nlm.nih.gov/gds)を例に実習します。

“This database stores curated gene expression DataSets, as well as original Series and Platform records in the Gene Expression Omnibus (GEO) repository. Enter search terms to locate experiments of interest. DataSet records contain additional resources including cluster tools and differential expression queries.” (GEOホームページより)

- 使いたいデータセットはダウンロードすることが可能
- 大規模なデータセットなので、「圧縮」されている

圧縮・解凍でつまづかないために

- OSによって使い分けられている(OSって…)
- Mac vs. Windows; and Linux
- Mac と Linuxは同類、Windowsは別もの
- 圧縮展開系のコマンドのまとめ (https://qiita.com/wnoguchi/items/cb0fa7c11b119e96f1e5)
- 色々な圧縮法があり、それに対応した解凍法がある
- 圧縮・解凍法は、拡張子で認識する
- 新規に圧縮・解凍法もできている
- 全部を覚える必要はない。必要に応じて、調べられれば十分
- 公共のデータベースでは、簡単なデータの確認・解析ができるようになっている場合もある

togotv “NCBI GEOのデータセットブラウザを使って公共データの遺伝子発現解析を行う 2019” (https://www.youtube.com/watch?v=-RdKozGzbbU)