DCB 統計処理法

TatsuoYAMAGUCHI

2022-04-17

はじめに

細胞生物学分野に所属する人間が最低限の統計処理を行うことができるように記した。そのため導出等無視して記載していくため、人によっては分かりづらい等あるかもしれない。そのときは以下の書籍等を参考にして欲しい。以下の書籍を参考文献であげておく。

統計学入門
自然科学のための統計学
東京大学教養学部統計学教室　東京大学出版会

心理統計学の基礎
続　心理統計学の基礎
南風原朝和著　有斐閣アルマ

多重比較法の基礎
永田靖・吉田道弘　著　サイエンティスト社

実験計画法入門
永田靖　著　日科技連

サンプルサイズの決め方**
永田靖　著　朝倉書店

データ解析のための統計モデリング入門
久保拓弥　著　岩波書店

実験計画

実験計画

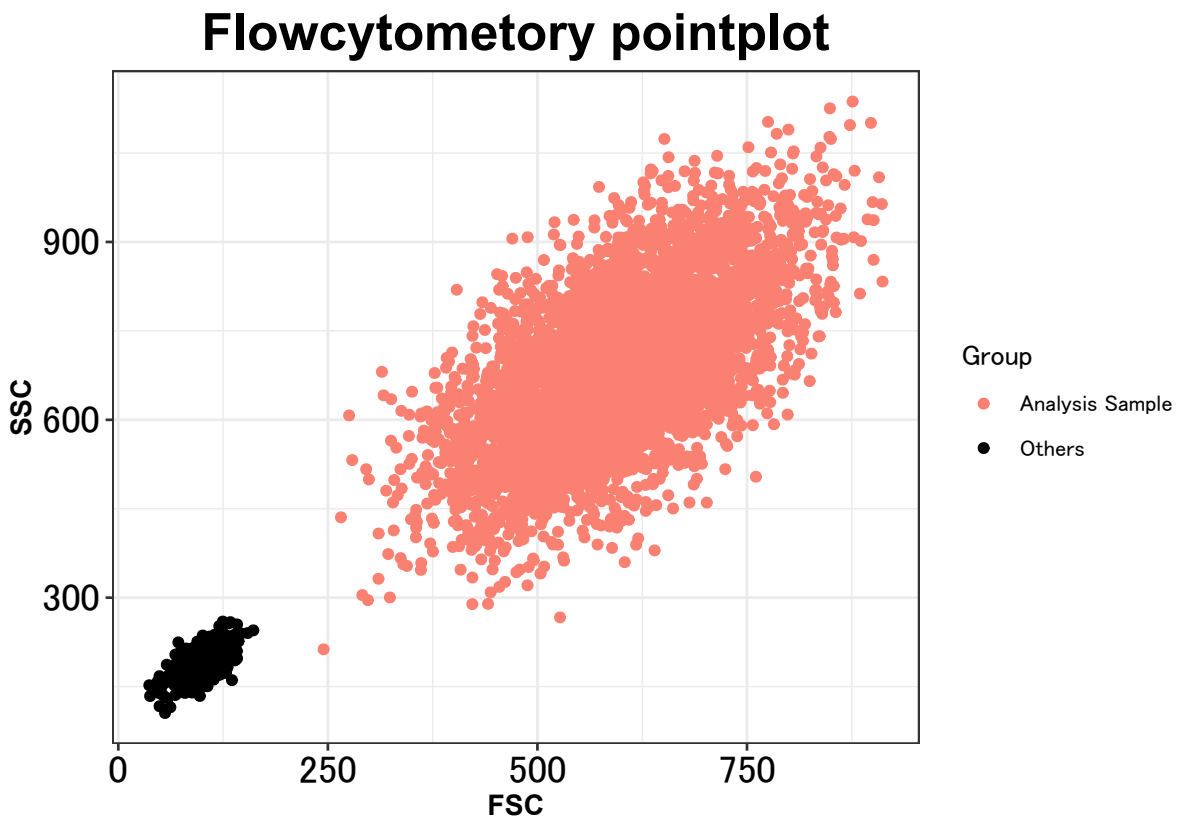
誤差伝搬

データの要約

データが出たらまずはグラフの描画である。ここではフローサイトメトリーを例に説明しよう。例えばフローサイトメトリーと呼ばれる細胞 1 個あたりの蛍光強度を測定する方法がある。ここでは細胞の膜タンパク質の量を蛍光抗体を用いて定量化すると仮定する。

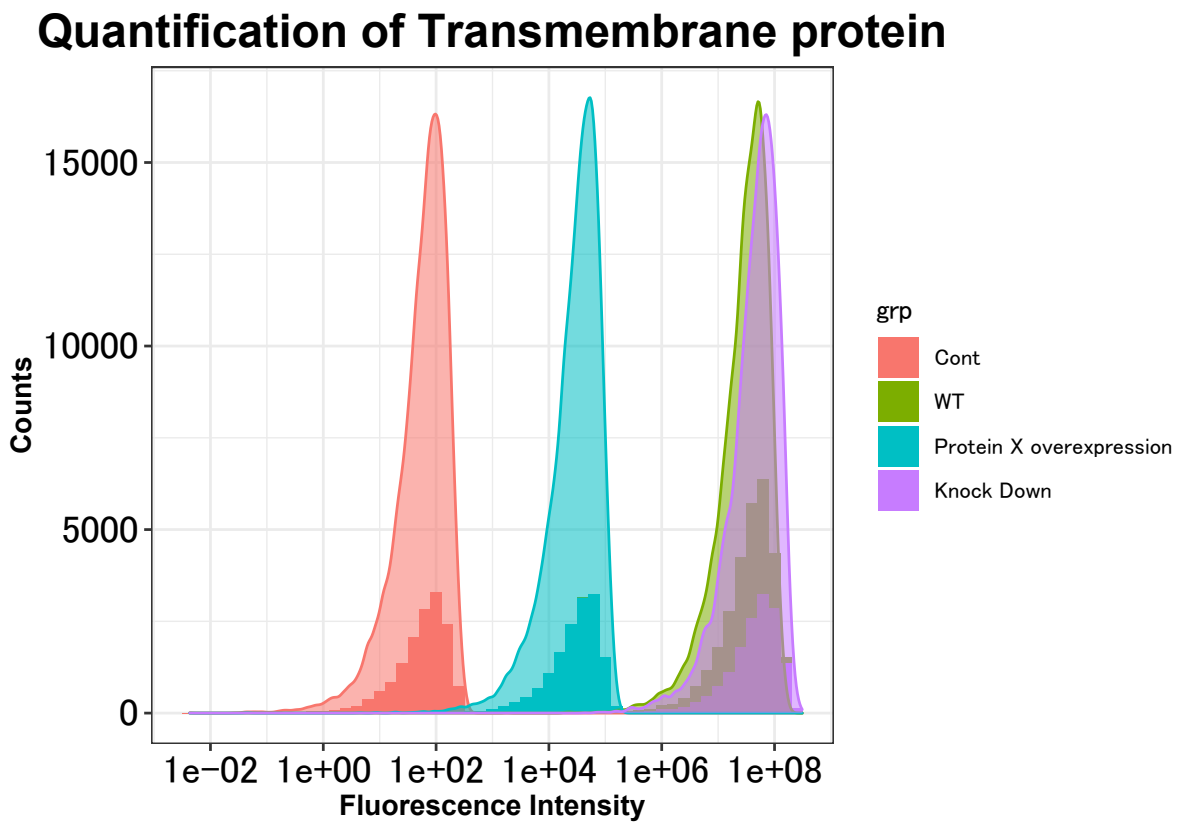
この仮想実験では、あるタンパク質 X を導入した細胞株、X を knock down した細胞株、未処理の細胞株と比較した。

まず測定時に x 軸（FSC（前面散乱光））と y 軸（SSC（側面散乱光））から測定したい細胞の範囲を決める。このグラフのことを散布図もしくは point plot と呼ぶ。



このサーモン色の範囲の細胞の膜タンパク質の量を測定すると、ヒストグラムと呼ばれる頻度を表すグラフを出力できる。

今回の仮想実験の結果から以下のようなヒストグラムを得ることが出来た。



このグラフは x 軸が蛍光強度、y 軸が対数蛍光強度の頻度を表しており、各色は各細胞株を示している。この結果から、(※実際の検出では 1.0×10^6 以上で測定することはない)

今回サンプルサイズ 7 で実験を行い、それぞれの各細胞株の蛍光強度を数値でまとめると以下のような結果が得られた。

```
## `summarise()` has grouped output by 'group'. You can override using the
## `.groups` argument.
```

group	number	n	mean	geo_mean	median	sd	CV
WT	1	15000	5286731318	33035648122	34153293962	30312866987	0.7521897
WT	2	15000	74944842411	33136951783	33968345182	29930106115	0.7512204
WT	3	15000	113104555044	28052014984	33615664609	29940062563	0.7530024
WT	4	15000	114866262195	8532235353	34491666150	30047657986	0.7468636
WT	5	15000	74083961814	18531961943	33792849785	30249076169	0.7590760
WT	6	15000	113498265458	17967126066	33904058920	30464584383	0.7575796
WT	7	15000	101992678704	5344949917	34701158647	30150219671	0.7440022

group	number	n	mean	geo_mean	median	sd	CV
Over expression	1	15000	20751514	31830674.66	33382826	30350505	0.7614692
Over expression	2	15000	70431146	1231519.54	33021387	29954281	0.7568548
Over expression	3	15000	78004985	53989985.95	33511234	30457124	0.7642261
Over expression	4	15000	64219705	59244.42	33695986	30160848	0.7563831
Over expression	5	15000	36586217	64624325.73	34207954	30223547	0.7528679
Over expression	6	15000	95742501	60717831.94	34032307	30332984	0.7592394
Over expression	7	15000	40357591	7633049.98	34358920	30062691	0.7499754

group	number	n	mean	geo_mean	median	sd	CV
Knock down	1	15000	240210023987	25357982046	51086413085	45365464306	0.7536942
Knock down	2	15000	77299184629	66873668338	51177670743	45145770782	0.7546393
Knock down	3	15000	189392456533	34680421808	50963921235	44928681222	0.7529941
Knock down	4	15000	45392404647	32651412195	50100902870	44887609933	0.7572284
Knock down	5	15000	71059046087	51489071415	49805585491	44966598984	0.7572404
Knock down	6	15000	135485834888	90469022246	49675965042	45267043696	0.7578109
Knock down	7	15000	101537490603	36599016380	49818335689	45261501188	0.7614641

この図で表されるのは number が本実験における replicate、つまりサンプルサイズ、mean が算術平均 (arithmetic mean)、geo mean が幾何平均 (geometric mean)、sd が標準偏差 (standard deviation)、CV が変動係数（標準偏差を算術平均で割った値）である。

幾何平均は、

$$Geometricmean = \sqrt[n]{x_1 \times x_2 \cdots x_n} = (\Pi x_i)^{1/n} = \exp(\frac{1}{n}\Sigma \ln(x_i))$$

である。ここで対数正規分布について説明する。対数正規分布は

$$f(x) = \frac{1}{\sqrt{2\pi\sigma}x} \exp(\frac{\ln(x) - \mu}{2\sigma^2})$$

で表される分布であり、平均値が

$$E(x) = \exp(+\frac{\sigma^2}{2})$$

分散が

$$V(x) = \exp(2 + \sigma^2)(\exp(\sigma^2) - 1)$$

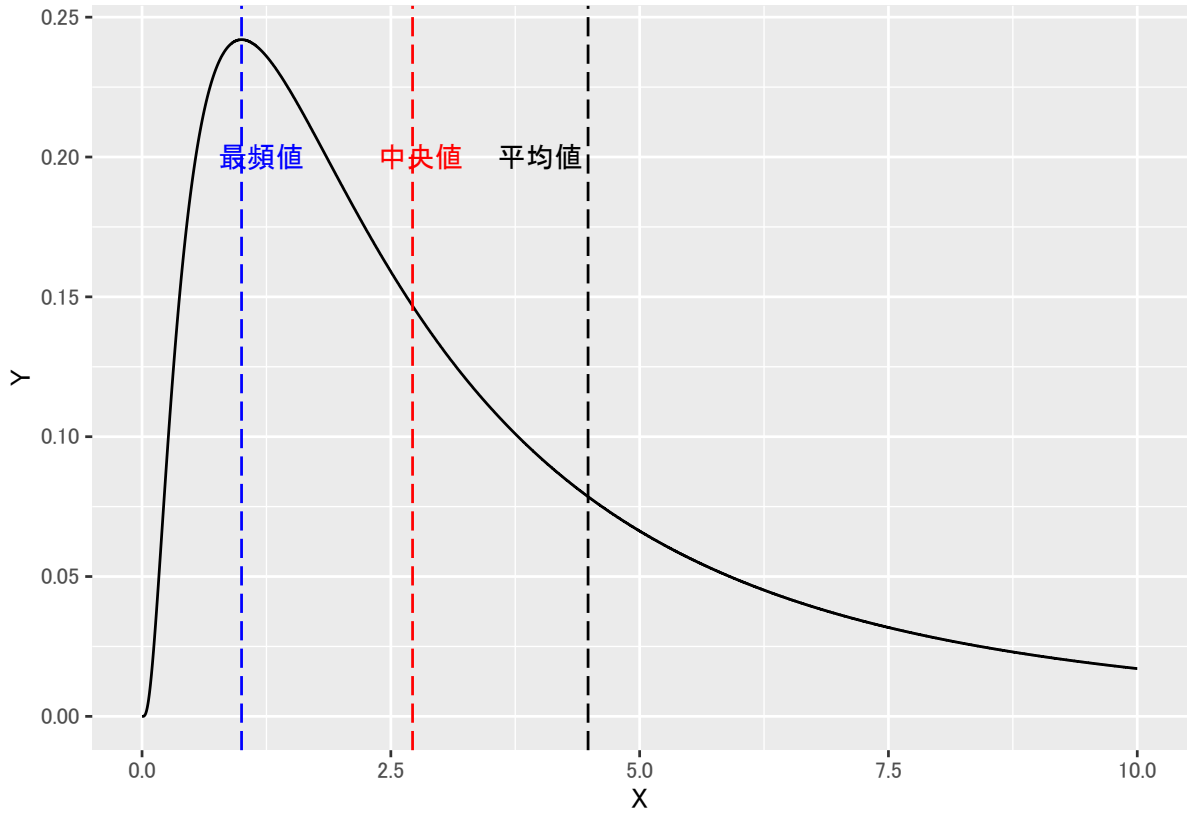
中央値が

$$Median = \exp(\mu)$$

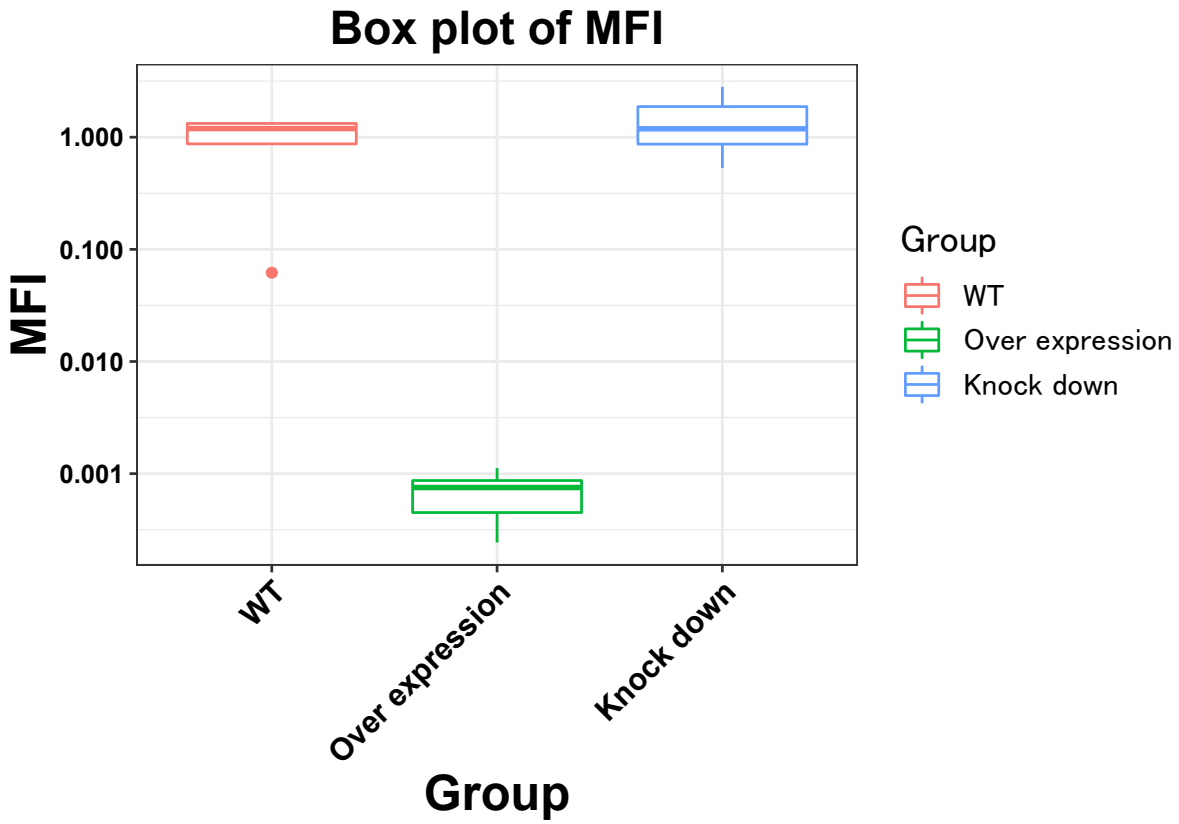
最頻値が

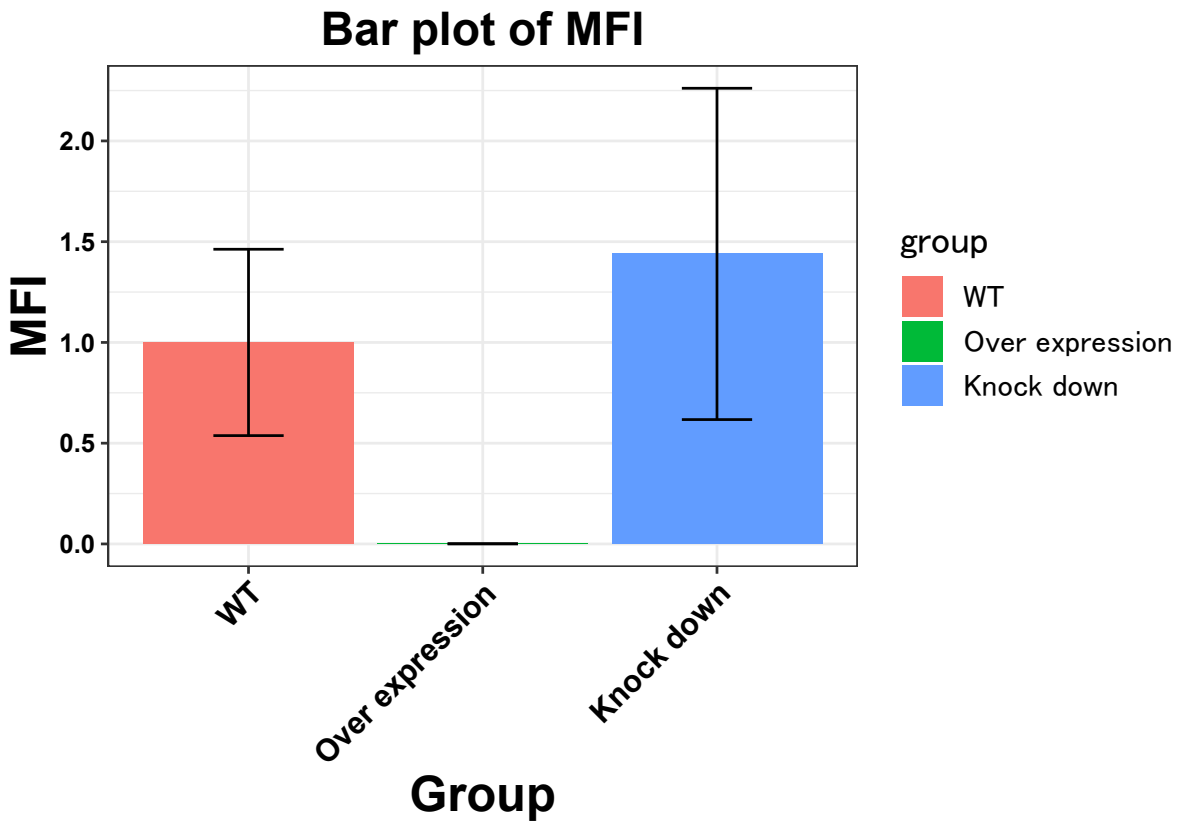
$$Mode = \exp(\mu - \sigma^2)$$

と表される。

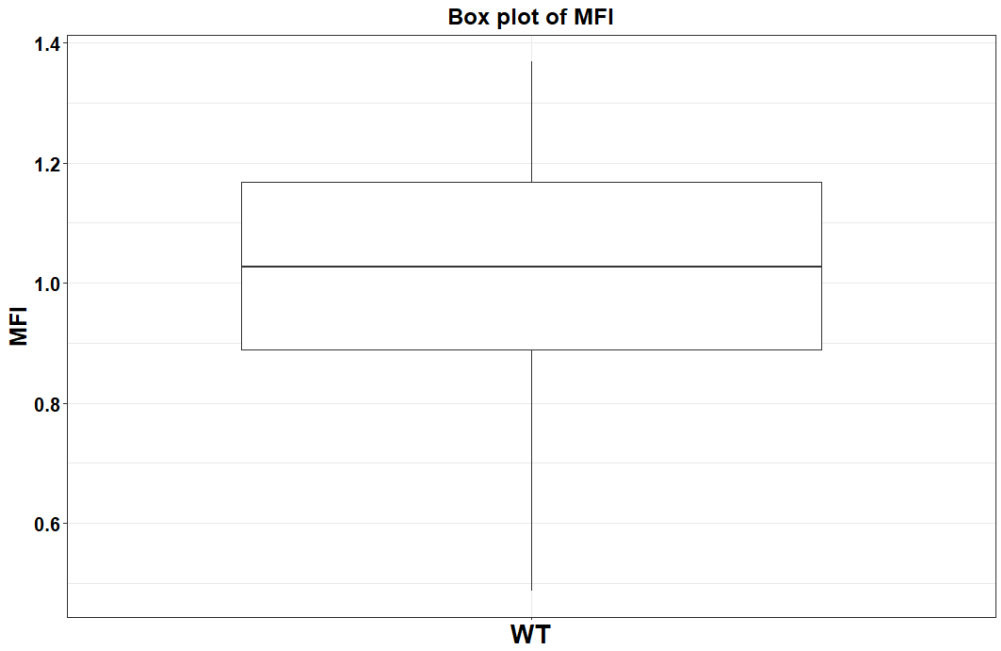


上記の数値ではわかりにくいので、WT で除したときの比率でグラフを描画してみる。そうすると以下のようになる。


Box plot of MFI
MFI
1.000
0.100
0.010
0.001
WT
Over expression
Knock down
Group
Group
WT
Over expression
Knock down


Bar plot of MFI
MFI
2.0
1.5
1.0
0.5
0.0
WT
Over expression
Knock down
Group
group
WT
Over expression
Knock down

上のグラフを箱ひげ図（Box plot)、下のグラフを棒グラフ (Bar plot) と呼ぶ。このグラフだと箱ひげ図が見づらいので、WT を例に箱ひげ図について説明する。



Box plot の上側のひげが最大値、Box の上側が 75 (%)、中心の横線が 50(%)（Median)、Box の下側が 25 (%)、下側のひげが最小値である。

分布

検定

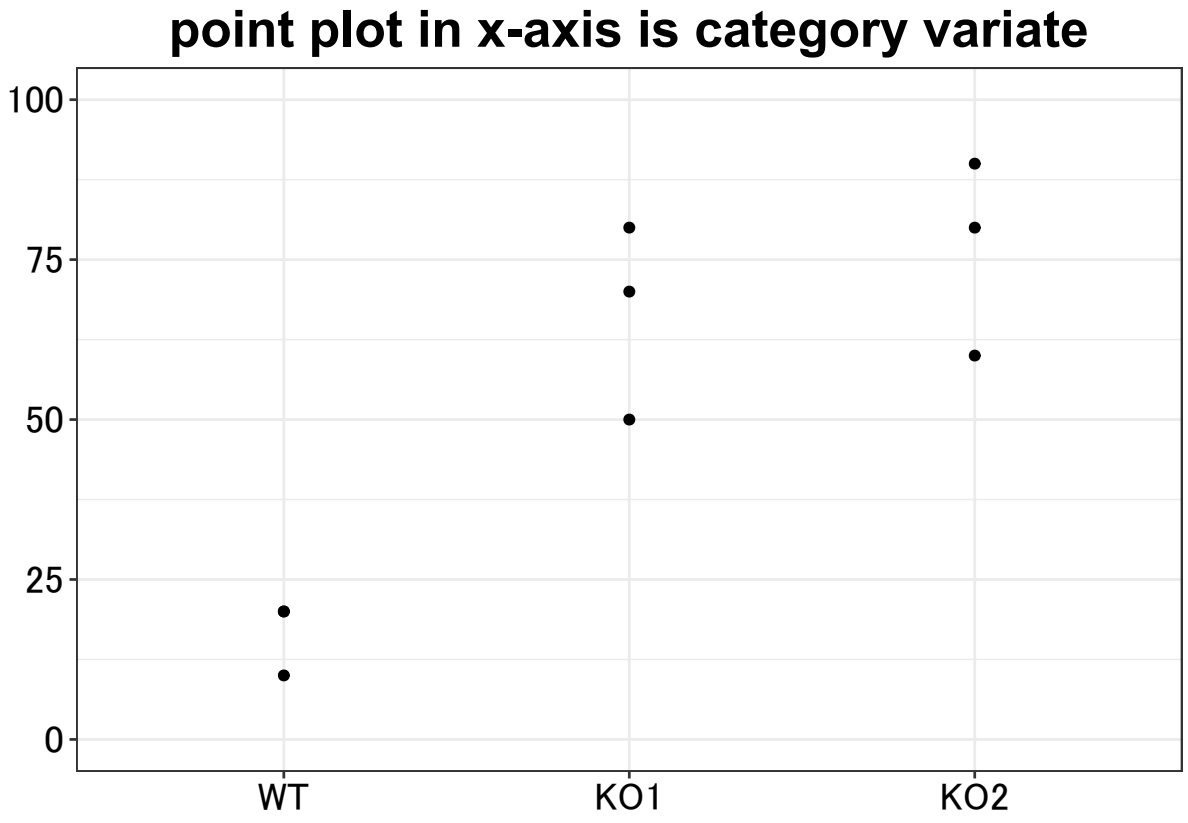
ではここで先ほどのデータの WT と Overexpression のあるタンパク質 X の mRNA 発現量を確かめるために、RT-qPCR をおこなった。その結果が以下の結果である。

このとき、実験で取り出してきたデータが母集団から取り出してきた、代表するデータであるならば、この平均値の推定を抽出する大元の母集団の平均に違いが見られるかを判別する、これが統計学的検定である。しかし統計学的検定で有意な差が見られても、生物学の観点から考えて有意でない場合も有るし、逆に統計学検定が有意でなくても生物学の観点から考えて有意な場合もある。こうした場合をできるだけ減らすために、予め実験をおこなう際にサンプルサイズ（n をいくつ取るのか）を決めたり、検出力の高い検定方法を選択したりする必要がある。とは言え、最終的には我々の持つ生物学的知見を大事にするべきであると私は思う。

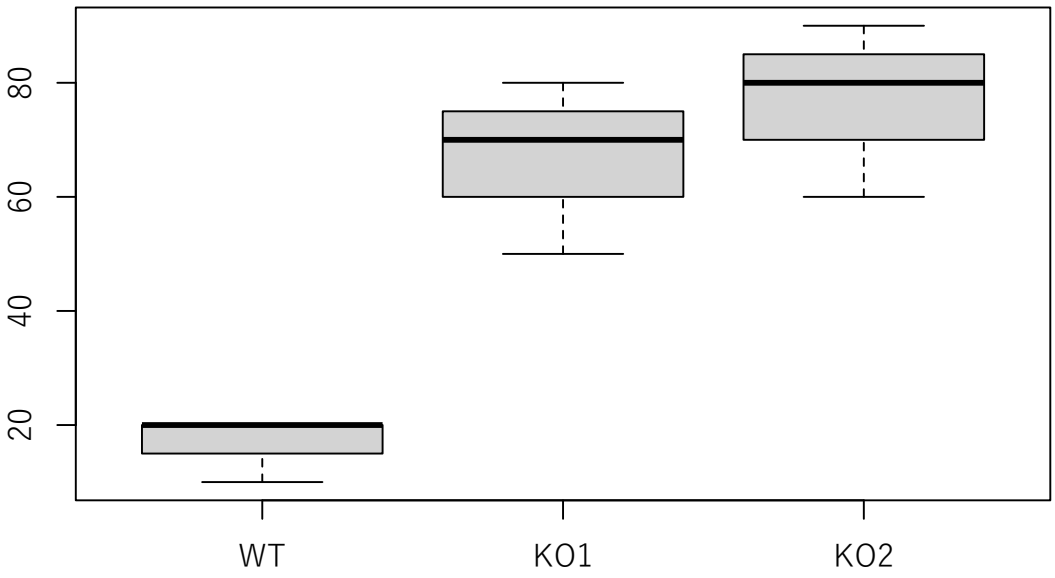
さて、

method	t_statistics	df	p-value	alternative
Two Sample t-test	-1.23221975699043	12	0.241465515786257	two.sided

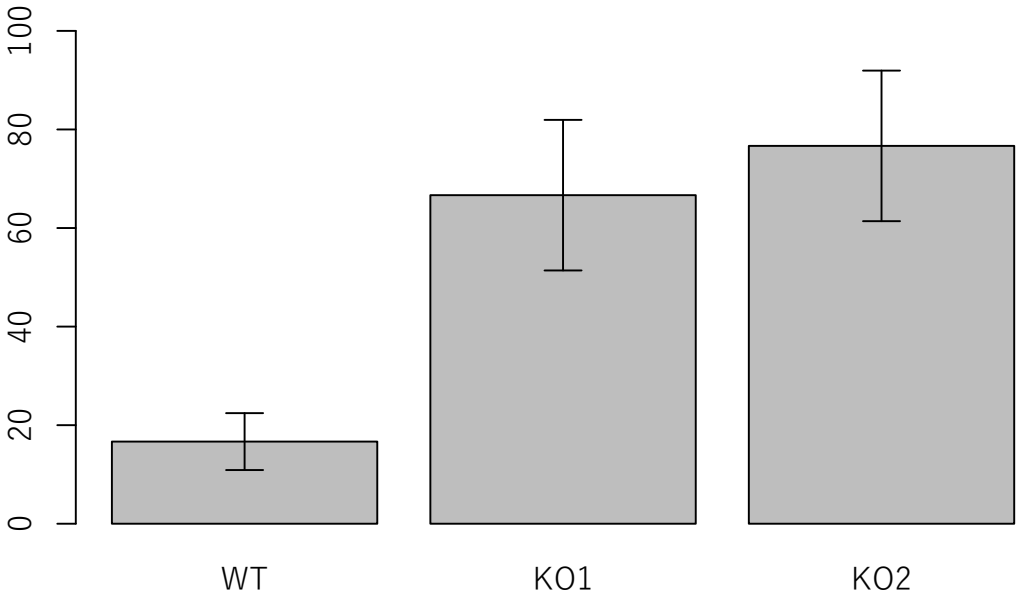
Two Sample t-test	-1.23221975699043	12	0.120732757893128	less
Welch Two Sample t-test	-1.23221975699043	9.45115943957309	0.247641420274116	two.sided
Welch Two Sample t-test	-1.23221975699043	9.45115943957309	0.123820710137058	less

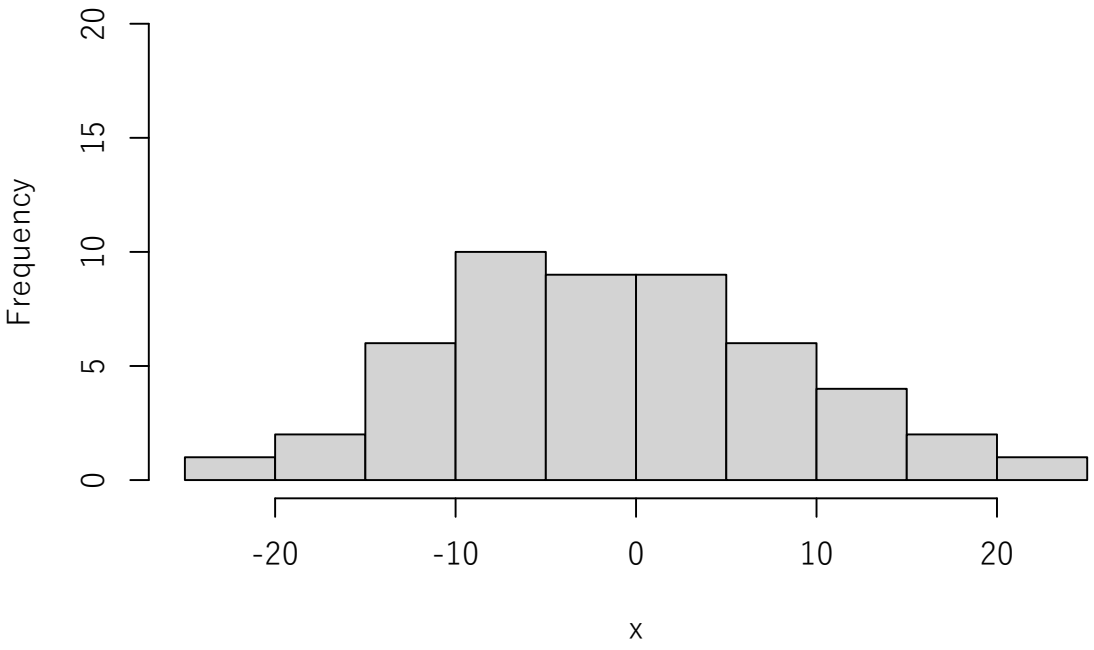


箱ひげ図



棒グラフ





検定

二群間の検定

対応の無い検定

対応のある検定

三群以上の処理

分散分析

推定

線形回帰モデル

単回帰分析

重回帰分析

非線形回帰モデル

EC_{50} IC_{50} の推定