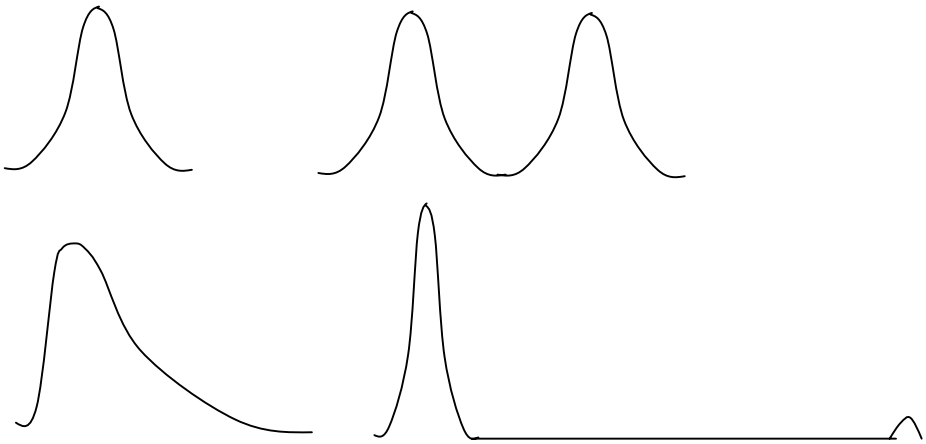
生物統計学の基礎知識１

サンプルの分布と検定

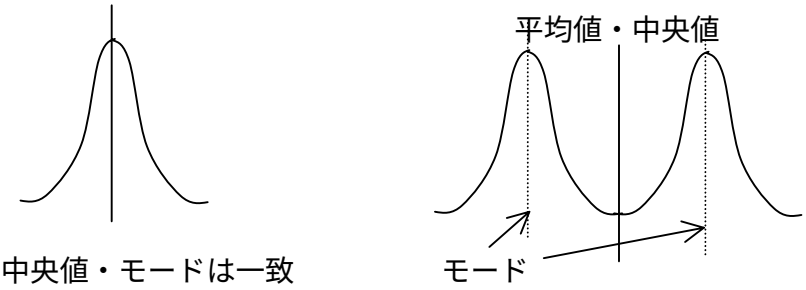
この章では生物統計学で用いられる古典的な手法を簡単に紹介します。実験レベルでは有用ですが、最近このような手法を用いて臨床研究が行なわれることは少なくなっています。

平均値、中央値、モードの違い

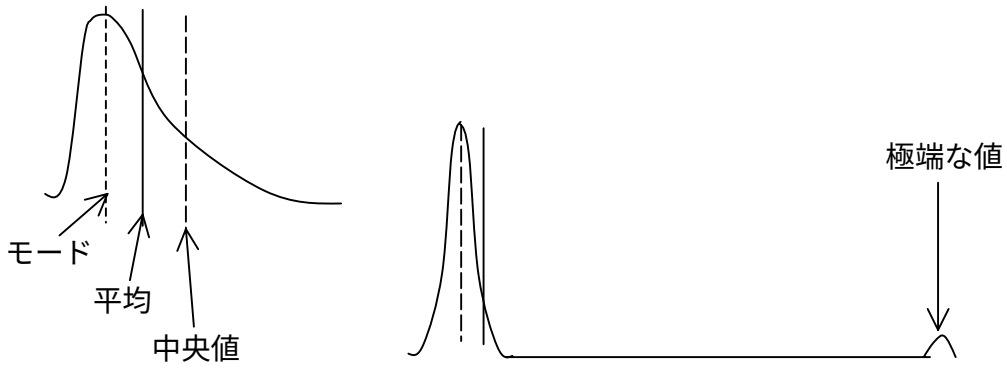
下のグラフに平均値、中央値、モードをそれぞれあてはめてみてください。



＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊

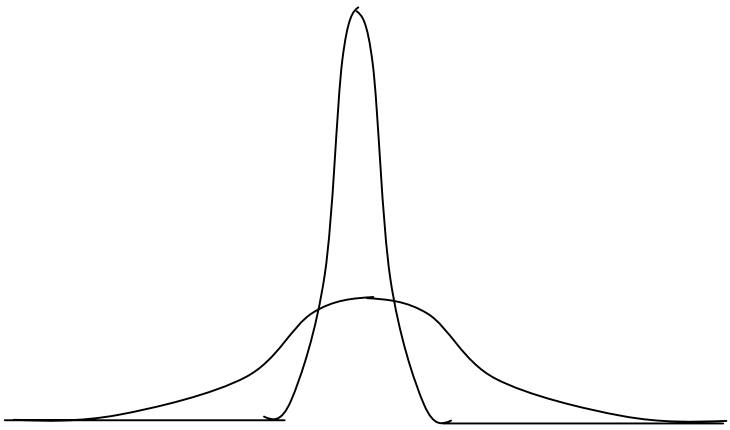


平均・中央値・モードは一致



中央値に比べ平均値は極端な値の影響を受け易い。

下の 2 つの分布を示す曲線は下の何によって識別されますか？
a. 平均値，b. 中央値、c. モード、d. 分散



上の２つの曲線において平均値、中央値、モードは一致しています。２つの分布を識別するのに必要なものは分散です。２つの分布で同じ平均値を持っていても、standard deviation が大きければ裾野の広い山型になるでしょうし、それが小さければ尖った形になります。

平均からの広がり（分散）を示すパラメータはvariance (s^2)であり、standard deviation (s) です。
つまり SD = $\sqrt{\text{var}}$ です。

$s^2 = 1/(n-1)\sum(x_i - \overline{x})^2$
$\overline{x}$ = mean, s = variance

以下の値の平均値と standard deviation を計算してください。
0.65, 0.80, 0.5, 0.35, 0.20, 0.13, 1.10, 0.70, 0.27, 0.05, 1.07, 0.10, 0.43

STATA を用いると容易に解答を得ることができます。

```
. sum

Variable |     Obs       Mean   Std. Dev.      Min       Max
---------+--------------------------------------------------
    var1 |      13    .4923077   .3545926       .05       1.1
```

答え：0.49±0.35

Binomial Distribution

Yes/No で示されるような事象、例えばコインを投げたとき表か裏か、男か女か、生きるか死ぬかなどはBernoulli random variable として知られます。例えば、ある街の喫煙者の確率は 0.29 であるとします。そして街行く人に声をかけてその人が非喫煙者である確率は 0.71 です。次の人も非喫煙者である確率は（0.71）^{2}です。それでは 2 人に声をかけて 1 人が喫煙者、もう 1 人が非喫煙者である確率はどうでしょうか？最初が喫煙者の場合 0.21 x 0.71 ですし、最初が非喫煙者の場合 0.71 x 0.29 です。

適当に 3 人の人を選んだときの喫煙者・非喫煙者の組み合わせとその確率は以下のようになります。

3 人非喫煙者	$0.71^3 \times {}_3C_3 = 0.3579$
2 人非喫煙者、1 人喫煙者	$0.71^2 \times 0.29 \times {}_3C_2 = 0.4386$
1 人非喫煙者、2 人喫煙者	$0.71 \times 0.29^2 \times {}_3C_1 = 0.1791$
3 人喫煙者	$0.29^3 \times {}_3C_0 = 0.244$
合計	1

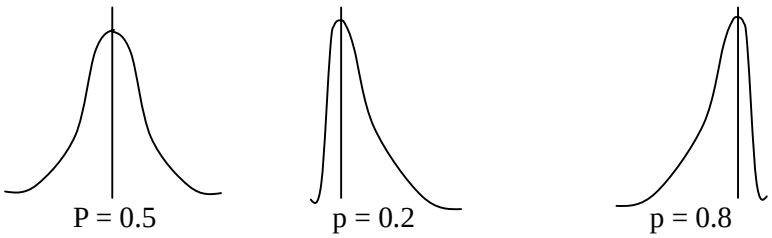
もしも Bernoulli 型の互いに独立したｎ回の実験をして、ｐの確率で成功するとしますと、変数がｘである確率は以下のように表されます。

$P(X=x) = nCx\ p^x(1-p)^{n-x}$

$Mean = np = 0.29 \times 10 = 2.9$

$SD = \sqrt{np(1-p)} = \sqrt{2.059} = 1.4$

P が 0.5 の時、結果はどちらが大きくなるかわからないので、SD は最大になり、逆に 0 または 1 近付いたとき、例えばほとんどの人が非喫煙者であるなど、SD は最小となります。



街角で喫煙の有無を 10 人の人に聞いたとします。喫煙者の確率は 0.29 ですから予想としては 10 人中 3 人より「吸ってるよ」という答えがかえってくることが期待されますが、確率的に誰も「吸っている」と答える人がいない可能性もありますし、逆に 10 人

とも「吸っている」ということもあるかもしれません。また 10 人中 5 人が吸っていると答える可能性は十分ありそうです。実際のところどうなっているでしょうか？上の公式を用いて STATA に計算させることができます。

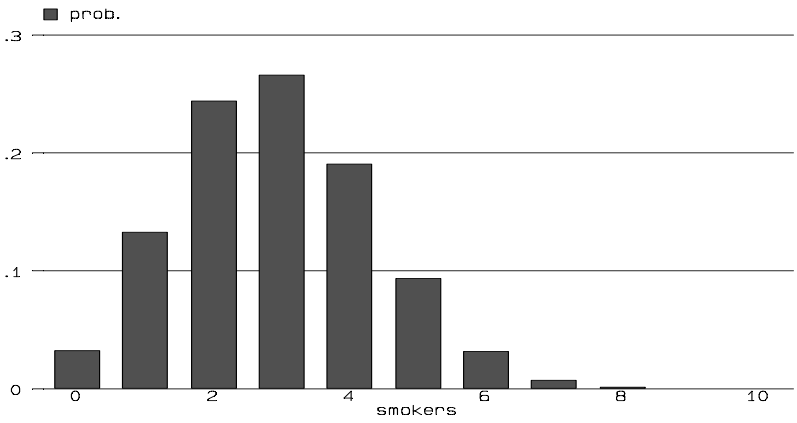
```
tablesq B 10 0 0.29

B(10,0.29) = 0
      Pr(k == 0) = 0.0326
      Pr(k >= 0) = 1.0000
      Pr(k <= 0) = 0.0326

. tablesq B 10 1 0.29

B(10,0.29) = 1
      Pr(k == 1) = 0.1330
      Pr(k >= 1) = 0.9674
      Pr(k <= 1) = 0.1655

. tablesq B 10 2 0.29

B(10,0.29) = 2
      Pr(k == 2) = 0.2444
      Pr(k >= 2) = 0.8345
      Pr(k <= 2) = 0.4099

. tablesq B 10 3 0.29

B(10,0.29) = 3
      Pr(k == 3) = 0.2662
      Pr(k >= 3) = 0.5901
      Pr(k <= 3) = 0.6761

. tablesq B 10 4 0.29

B(10,0.29) = 4
```

```
      Pr(k == 4) = 0.1903
      Pr(k >= 4) = 0.3239
      Pr(k <= 4) = 0.8663

. tablesq B 10 5 0.29

B(10,0.29) = 5
      Pr(k == 5) = 0.0933
      Pr(k >= 5) = 0.1337
      Pr(k <= 5) = 0.9596

. tablesq B 10 6 0.29

B(10,0.29) = 6
      Pr(k == 6) = 0.0317
      Pr(k >= 6) = 0.0404
      Pr(k <= 6) = 0.9913

. tablesq B 10 7 0.29

B(10,0.29) = 7
      Pr(k == 7) = 0.0074
      Pr(k >= 7) = 0.0087
      Pr(k <= 7) = 0.9988

. tablesq B 10 8 0.29

B(10,0.29) = 8
      Pr(k == 8) = 0.0011
      Pr(k >= 8) = 0.0012
      Pr(k <= 8) = 0.9999

. tablesq B 10 9 0.29

B(10,0.29) = 9
      Pr(k == 9) = 0.0001
```

```
      Pr(k >= 9) = 0.0001
      Pr(k <= 9) = 1.0000

. tablesq B 10 10 0.29

B(10,0.29) = 10
      Pr(k == 10) = 0.0000
      Pr(k >= 10) = 0.0000
      Pr(k <= 10) = 1.0000
```

これをグラフにすると、

.



３をピークに向かって左の方にやや偏った(skewed)グラフとなります。

ある研究者は両親が慢性気管支炎をもっている場合、子供が生後 1 年以内に気管支炎になるのは 20 人に 3 人の割合であることを発見しました。一般的には乳児期気管支炎を発症する率は 5%だとします。これは大きな発見でしょうか、それとも偶然でしょうか？ 20 人のうち少なくとも 3 人が気管支炎である確率はどれくらいですか？

$_{20}C_K\,(0.05)^k(0.95)^{20-k}$, K = 0, 1, 2, ... 20

少なくとも 3 人が気管支炎ということは、0, 1, 2, 人が気管支炎である確率を最初に求め、１から引いたほうが楽です。

$_{20}C_0\,(0.05)^0(0.95)^{20} = 0.3585$

$_{20}C_1\,(0.05)^1(0.95)^{19} = 0.3774$

$_{20}C_2\,(0.05)^2(0.95)^{18} = 0.1887$

1 – (0.3585 + 0.3774 + 0.1887) = 0.0754

つまり 20 人のうち少なくとも 3 人が気管支炎である可能性は 7.5%であり、統計学においえる一般的 cut off は 5%ですから、この例題は”たまたま”と考えた方がよさそうです。ここで”少なくとも 3 人”でなく、”3 人”に設定すると確率が低くでてしまい結果を間違って推論してしまうことがありますので注意してください。

例えば疾患 X の頻度は 0.00001 です。2,500,000 人の集団を対象に調査したところ 36 人疾患 X が見つかりました。これは偶然でしょうかそれとも統計学的に有意に多いのでしょうか？

とても手で計算するわけにはいかないので、STATA を使います。

```
. bitesti 2500000 36 0.00001

             N      Observed k      Expected k      Assumed p      Observed p
------------------------------------------------------------------------------
       2500000              36              25        0.00001        0.00001

  Pr(k >= 36)                 = 0.022458    (one-sided test)
  Pr(k <= 36)                 = 0.985448    (one-sided test)
  Pr(k <= 14 or k >= 36) = 0.034859    (two-sided test)
```

0.05 より Pr(k >= 36)が小さいので偶然ではなさそうです。少し本格的に調査するべきかもしれません。

結核の病状を把握するために頻回にレントゲン写真を撮影したことが乳癌発生と関係あるかどうかを調査し、以下のような結果を得ました。

	暴露	非暴露	合計
乳癌	41	15	56
Person-years	28,010	19,017	47,027

もしも乳癌の発生がレントゲン施行と関係ないとすれば、28,010/47,027 と同じ比率で乳癌の発生を認めるはずです。そこで p = 28,010/47,027 として計算します。

```
. bitesti 56 41 28010/47027

             N      Observed k      Expected k      Assumed p      Observed p
```

56	41	33.35446	0.59562	0.73214

Pr(k >= 41) = 0.023830 (one-sided test)
Pr(k <= 41) = 0.988373 (one-sided test)
Pr(k <= 25 or k >= 41) = 0.040852 (two-sided test)

Two-sided test においても有意差を認めております。つまり、かつて結核に対して頻回にレントゲン撮影を行なったことが一部の人で乳癌を引き起こすきっかけになっていた可能性を示唆しています。

疫学調査をもう少し勉強してみましょう。
pのmaximal likelihood estimate はA/Nであり、そのvariance はA(N – A)/N^3 です(Aは観察された疾患数であり、Nはcohort sizeです)。

Doll&Hill による British physicians study を例に考えてみましょう。健康な男性医師を何らかの疾患が発生しないかどうか 4 年 5 ヶ月経過観察しました。そして喫煙者、非喫煙者で疾患の発生に違いがあるかどうか比べています。

	喫煙者	非喫煙者	合計
死亡	1582	166	1748
健康	27116	5630	32746
合計	28698	5796	34494

そこでindependence and homogeneity assumption が成り立つものとして、上で説明したbinomial distribution の考え方を用いて、喫煙者、非喫煙者それぞれのcohort における死亡者(case)のvariability について検討します。喫煙者におけるpのmaximal likelihood estimate はA/N=1582/28698 = 0.0551 であり、そのvariance はA(N – A)/N^3 = 1582 x 27116/$(28698)^3$です。95% CI = p ± 1.96 √ var = (0.0243, 0.0329)となります。非喫煙者におけるpのmaximal likelihood estimate はA/N=166/5796 = 0.0532 であり、そのvariance はA(N – A)/N^3 = 166 x 5630/$(5796)^3$です。95% CI = p ± 1.96 √ var = (0.0243, 0.0329)となります。

薬剤のスクリーニング

薬剤を開発する最初の過程で、非常に多くの物質がスクリーニングにかけられます。その際、なるべく少ないマウスの数でなるべく正確なデータを得たいと思います。１つの物質に対して多くのマウスを用いればより正確なデータが得られるのはあたりまえですが、命あるものを無益に殺すのは良くないことです。また何百という可能性のある物質から最も薬剤として可能性のあるものを選ぶのですから、１つのかけられる時間と費用も限られています。Binomial distribution の概念を用いて効率的に物質の可能性を探るにはどのようにしたらよいでしょうか？

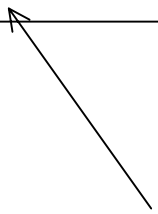
例えば避妊薬の新薬開発をしているとします。ある既存経口避妊薬をハムスターに投与し妊娠状態を確認したところ以下のデータを得ました。

	妊娠	非妊娠	合計	妊娠率(%)
コントロール	62	6	68	91
低用量	11	19	30	37
高用量	1	24	25	4

このデータを踏まえて経口避妊薬の新薬を数ある物質の中からスクリーニングしたいと思います。まずはスクリーニングなので各物質をそれぞれ 8 匹のハムスターに投与し 4 匹以下が非妊娠であれば active, 5 匹以上が妊娠すれば inactive としようと思います。この方法を採用したとき、ある物質の本当の妊娠率を 0.2 と想定したとき、active と判断する確率はどれくらいですか。本当の妊娠率を 0.0, 0.1, 0.2, . . . 0.9, 1.0 と変えていったとき accept する確率はどのように変化しますか？グラフで示してください。

N=8, p = 0.2

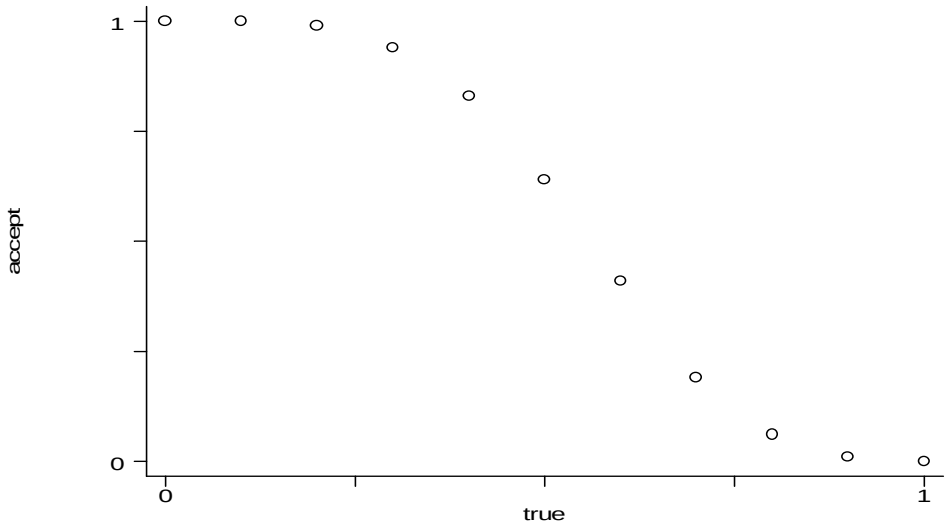
Pregnant	probability of accept	cumulative
0	0.1678	0.1678
1	0.3355	0.5033
2	0.2936	0.7969
3	0.1468	0.9437
4	0.0459	0.9896
5	0.0092	0.9988
6	0.0011	0.0000
7	0.0001	1.0000
8	0.0000	



スクリーニングしている物質の本当の避妊効果がP＝0.2であるときにその物質がスクリーニング検査において active であると宣言する確率　(0.99)

以下同様に行なうことのよって下の表とグラフを得る事ができます。

	P=0	P=0.1	P=0.2	P=0.3	P=0.4	P=0.5	P=0.6	P=0.7	P=0.8	P=0.9	P=1
0	1	0.43	0.17	0.06	0.02	0.00	0.00	0.00	0.00	0.00	0.00
1	0	0.81	0.50	0.26	0.11	04	0.01	0.00	0.00	0.00	0.00
2	0	0.96	0.80	0.55	0.32	0.14	0.05	0.01	0.00	0.00	0.00
3	0	0.99	0.94	0.81	0.59	0.36	0.17	0.06	0.01	0.00	0.00
4	0	1.00	0.99	0.94	0.83	0.64	0.41	0.19	0.06	0.01	0.00
5	0	1.00	1.00	0.99	0.95	0.86	0.68	0.45	0.20	0.04	0.00
6	0	1.00	1.00	1.00	0.99	0.96	0.89	0.74	0.50	0.19	0.00
7	0	1	1	1	1	1.00	0.98	0.94	0.83	0.57	0.00
8	0	1	1	1	1	1	1	1	1	1	1



横軸はそのスクリーニングにかけている物質を用いたときの真の妊娠率です。0 から 1 であることは間違いありませんが、「神ののみぞ知る」で誰も知りません。8 匹のうち 4 匹以下の妊娠であればこの物質が避妊薬になり得る可能性があるものと考え accept するルールに最初に設定しました。縦軸はそのルールに従って accept する確率です。例えば真の妊娠率が 0 や 0.1 であった場合、たった 8 匹でも確実に OK サインを出せます。0.2 や 0.3 の経口避妊薬としてよく効きそうな物質を捨ててしまうことはあまりなさそうです。しかし 0.4 あたりからあやしくなり、半数にしか有効でない物質が避妊薬として約 6 割 accept されてしまいます。もしも妊娠率を 0.2 以下に抑えられる物質でなければ絶対駄目ということであれば 8 匹中 3 匹以下の妊娠で accept するようにルールを変えるべきかもしれません。そうすれば 0.2 で 94%は accept し、0.5 のものは 36％しか accept しません。上のようなカーブを **Operating Characteristic Curve (OC)**と呼びます。よってスクリーニングの際、何匹中何匹が陽性（あるいは陰性）でその物質を次の検査にまわすかは検査をする人の思惑と OC カーブで決定します。さらにもう一度スクリーニングをかけることにより(two stage screening)、さらに精度を上げることができます。

Poisson Distribution

毎年交通事故に巻き込まれる確率は 0.00024 だとします。これは事故に遭うか、遭わないかなので binomial situation です。しかし、ｎが非常に大きくて、ｐが非常に小さい時、binomial distribution として計算するのは非常に大変です。しかもｐが 0 に近いので 1-p は 1 に近値します。そういう場合 Poisson 分布を用います。稀な事象をみる場合もそうですが、疫学調査でしばしば用いる person-time を用いると分母が分子に比べて相当大きくなるのでしばしば Poisson distribution の適応になります。

Poisson distributioin には、2 つの重要な仮定があります。１つは independence assumption （独立仮説）です。例えば伝染病流行のような場合には、B さんの感染症になる確率は一緒に働いている A さんが感染症になると変ってしまうからです。このような場合には、Poisson を用いることはできません。もう１つは Stationary assumption（静止仮説）で、結果発生は調査期間内一定でなくてはなりません。時間の経過とともに疾患発生頻度が上がれば Poisson 分布を用いる事はできません。例えば、白血病化学療法開始後の再発をみるとします。再発は 1 年の治療終了後 1 年間に多発していたとします。このような場合には Poisson よりは、時間的要素を含む Hazard model を用いるべきです。よって Poisson 分布を用いる場合には調査期間は短くするべきです。

$P(X=x) = e^{-\lambda} \lambda^{x} / x!$

ｘは 0 から無限大の整数で、λは平均、e=2.7182

もしもｐが 0 に近付くと、1－p は 1 に近付くため、平均および variance は np となります。

毎年交通事故に巻き込まれる確率は 0.00024 だとします。今年 1 万人当たり 4 人が交通事故に遭う確率は？

$\lambda = np = 10{,}000 \times 0.00024 = 2.4$

$P(X=4) = e^{-2.4} (2.4)^4 / 4! = 0.1254$

答えは 12.4%ということになります。

毎年ある村に白血病が 3 人発生するとします。もし何人になったら「今年は白血病が多発している、何かおかしいぞ」と警笛をならしますか？

$\lambda = np = 3$

$P(X=x) = (x - 3) / \sqrt{3} > 1.645 \ (p=0.05)$

X = 6

6 人を超えた警笛をならします。

先に述べた通り、Poisson distribution は分母に比べ分子が非常に小さい場合に用いるので、臨床研究においては分母を person-time にしているようなときに用います。

μ = (person time) x (incidence rate)

$P(X=x) = e^{-\mu}\mu^{x}/x!$

λ: expected number of events per unit time

μ: expected number of events over the time period t

$\mu = \lambda t$

μ の maximal likelihood of estimate(MLE) を A とすると、incidence rate (IR)に対する MLE は A/person-time (PT) となります。それでは binomial distribution の際用いた Doll & Hill のデータを再利用してみましょう。

	喫煙者	非喫煙者
死亡	1582	166
Person-years (PY)	123436	25250

喫煙者：MLE=1582/123436 = 0.0128

死亡者数　95% CI = $1582 \pm 1.96\sqrt{1582}$ = (1504, 1660),

incidence rate 95% CI = (1504/123436PY, 1660/123436PY) = (0.0122/PY, 0.0134/PY)

非喫煙者：95% CI = $166 \pm 1.96\sqrt{166}$ = (141, 191)

incidence rate 95% CI = (141/25250PY, 191/25250PY) = (0.0558/PY, 0.00756P/Y)

となります。

下の表はポリオ感染症による死亡例を示しています。

	1968	1969	1970	1971	1972	1973	1974	1975	1976
death	24	13	7	18	2	10	3	9	16

このデータの平均は 11.3 であり、variance は 51.5 です。Outbreak の年とそうでない年の gap が流行する感染症では存在しますから variance は大きくなって然るべきです。しかし Poisson distribution では mean と variance はμに等しくなりますから、これからしても outbreak を起こしうる感染症はいくら希だからといって Poisson を当てはめて計算することができない良い例です。

下の list は、９つの航空会社の傷害事故(injury)を調べた結果です。n は飛行回数当りの傷害事故数の割合です。飛行機事故は比較的希なことであり、Poisson distribution と考えられます。我々はある装置（XYZ）を持っている(1)ことが、事故につながったのではないかと考え検証してみようと思います。

. list

	airline	injuries	n	XYZowned
1.	1	11	.095	1
2.	2	7	.192	0
3.	3	7	.075	0
4.	4	19	.2078	0
5.	5	9	.1382	0
6.	6	4	.054	1
7.	7	3	.1292	0
8.	8	1	.0503	0
9.	9	3	.0629	1

. poisson injuries XYZowned, exposure(n) irr

Iteration 0: log likelihood = -23.027197
Iteration 1: log likelihood = -23.027177
Iteration 2: log likelihood = -23.027177

Poisson regression Number of obs = 9
LR chi2(1) = 1.77
Prob > chi2 = 0.1836
Log likelihood = -23.027177 Pseudo R2 = 0.0370

injuries	IRR	Std. Err.	z	P>|z|	[95% Conf. Interval]	
XYZowned	1.463467	.406872	1.370	0.171	.8486578	2.523675
n (exposure)						

. gen lnN=ln(n)

XYZ を装備している飛行機は装備していない飛行機と比較して 1.46 倍事故を起こしやすいのですが、P = 0.171 であり、95%CI も１を含んでおり有意ではありません。つまり、「上のデータからは XYZ を持っていることが飛行機事故とつながるとは言えない」と結論できます。

先の例では incidence rate ratio を考えましたが、次に

$$\text{rate} = e^{\beta o + \beta 1 XYZowned}$$

よって観察された数は

$$\text{count} = n\, e^{\beta o + \beta 1 XYZowned} = e^{\ln(n) + \beta o + \beta 1 XYZowned}$$

です。

. poisson injuries XYZowned lnN

Iteration 0: log likelihood = -22.333875
Iteration 1: log likelihood = -22.332276
Iteration 2: log likelihood = -22.332276

Poisson regression Number of obs = 9
LR chi2(2) = 19.15
Prob > chi2 = 0.0001
Log likelihood = -22.332276 Pseudo R2 = 0.3001

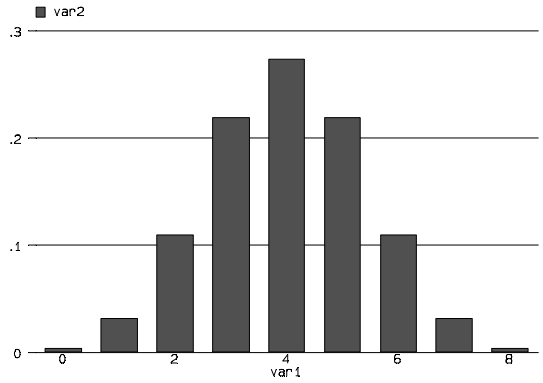
injuries	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
XYZowned	.6840667	.3895877	1.756	0.079	-.0795111	1.447645
lnN	1.424169	.3725155	3.823	0.000	.6940517	2.154285
_cons	4.863891	.7090501	6.860	0.000	3.474178	6.253603

$e^{0.684} = 1.98$

point estimate は 1.98 倍まで増えましたが、まだ有意な増加とはいえません。

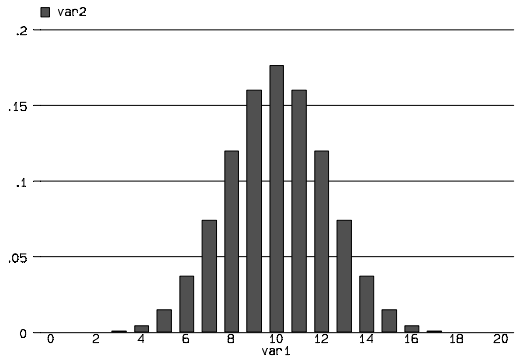
正規分布(normal distribution)

コインを 8 回投げてでた表の数を数えます。可能性として一度も表のでない状態から 8 回とも表のでることまでありえるわけですが、０から８までのそれぞれの確率はどうなるでしょうか。棒グラフでしめしてください。



確率分布を上のような形で示したものを probability distribution と呼びます。確率の総和は１となります。

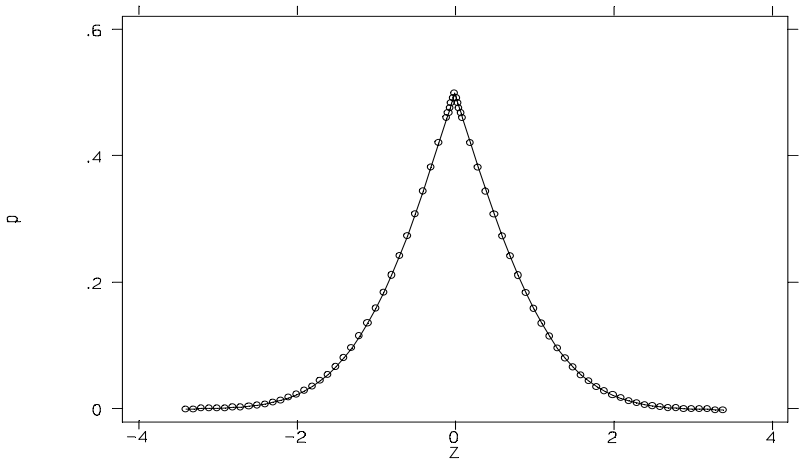
回数を増やして 8 回から 20 回にしたらどうなりますか？



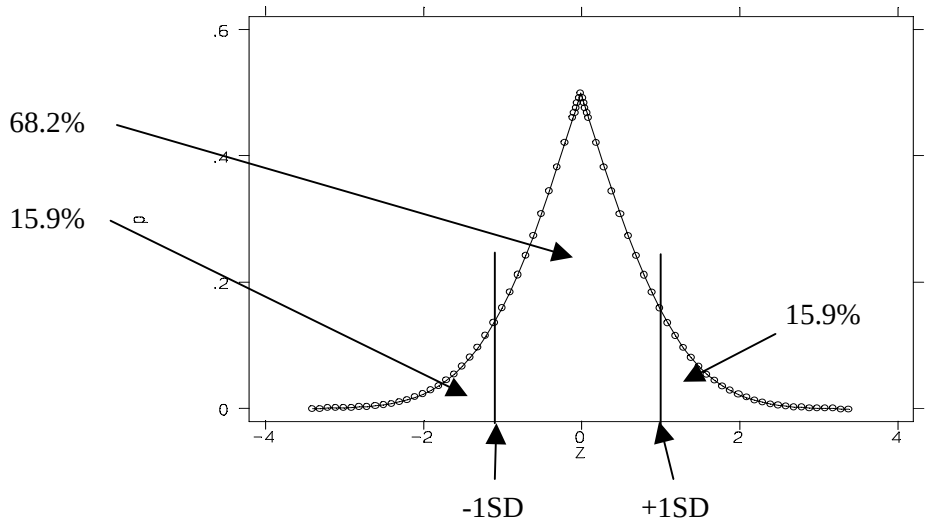
8 回よりも 20 回の方がカーブがなだらか（スムース）です。もしも施行回数を無限近付けたとすると完全な曲線となるはずです。このようにして描かれたスムースなカーブ（確率 p が一定で無限大行なったとき）を正規分布(normal distribution/Gaussian distribution/bell-shaped distribution)と呼びます。正規分布においては平均、中央値、モードは全て一致します。そして平均値μと standard deviation (σ)によってその形は規定されます。

臨床で用いられる多くのデータが正規分布をなします。たとえば身長、体重、血圧、血糖など数えたら限がありません。

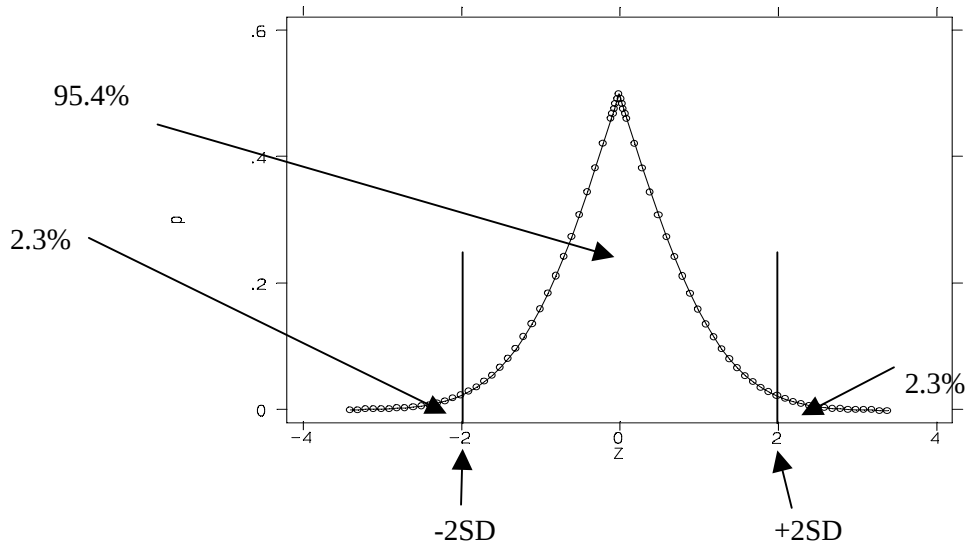
正規分布の中でも平均値μが 0 であり、standard deviation (SD) σ が 1 であるものを standard normal distribution とします。いわばスペードのエースのようなものです。



横軸 z は SD を示しています。

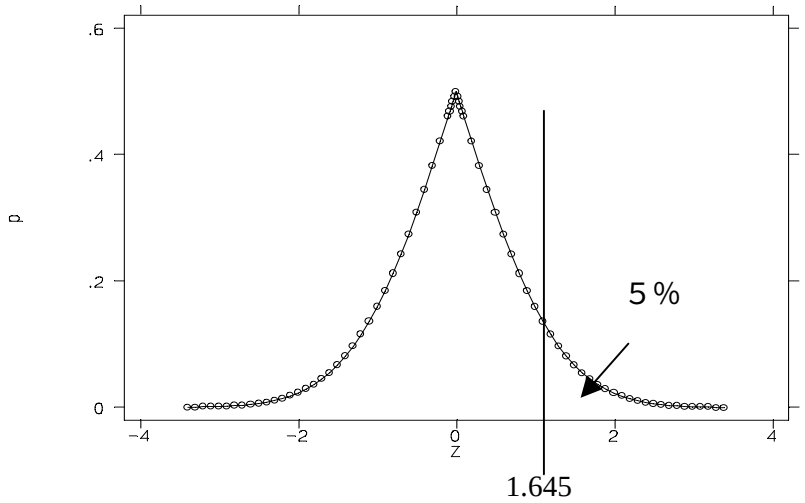


μ ± 1SD で囲まれる面積は全体の 68.2%を占め、よって残りの片側は 15.9%を占めます。

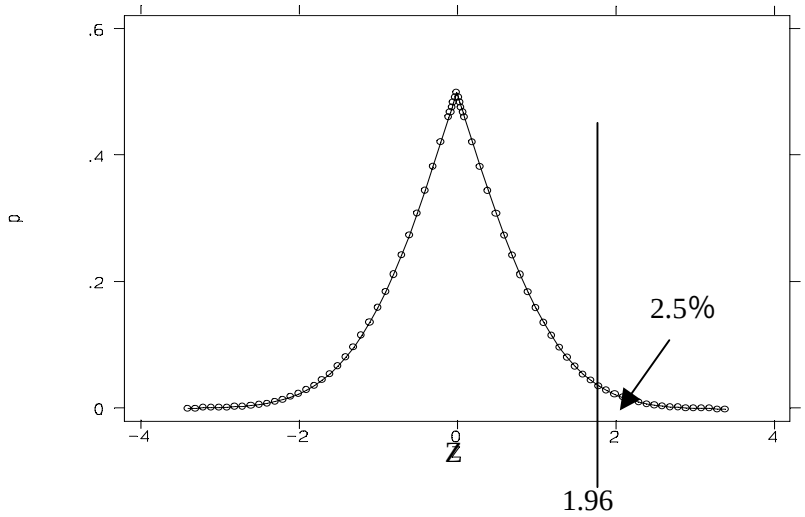


2SD で面積を区分した場合には上図のようになります。

上 2 つの図は横軸である z を中心に考えましたが、今度は逆に面積を中心に考えます。
例えば片側 5%の面積は Z にしていくつに相当しますか？



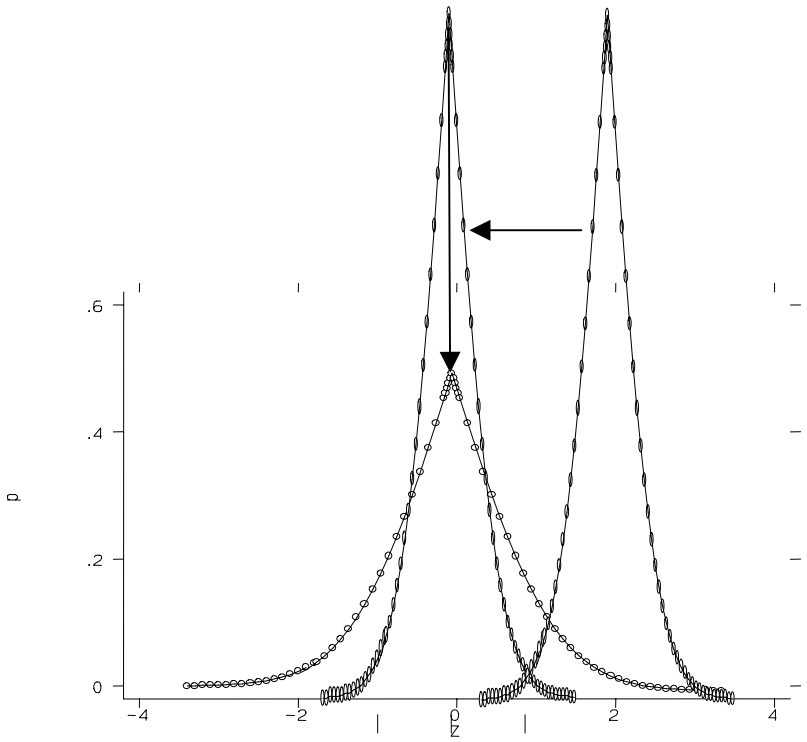
同様に片側面積 2.5%に相当する横軸 z はいくつですか？



この 1.645 と 1.96 はｐ値がそれぞれ片側検定、両側検定で 0.05 に相当する数値なので呪文のように覚えてしまった方が便利です。

Standard normal distribution curve を少し変形してみましょう。例えば平均値が 2.0、SD が 0.5 であるとします。Distribution はどのように変化しますか？

SD が半分になりますから standard よりスリムになりかつのっぽになります。平均は０から２ですから２平行移動します。



それでは上で新たに作った正規分布で X = 3.0 はもとの standard normal distribution においていくつ(Z)に相当するでしょうか？

新たに作った正規分布を standard normal distribution に戻す操作をすれば自ずと X=3.0 に相当する Z を求めることができます。3.0 より 2 左に平行移動し（マイナス２）、SD は 0.5 から 1.0 になるので倍の幅広になります。具体的に数式で示すと、

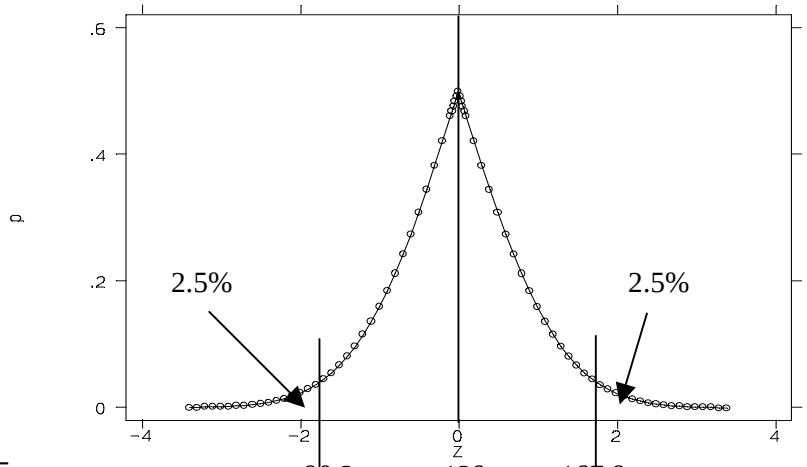
Z = (X – 2)/0.5　　すなわち Z = (3 – 2)/0.5 = 2 に相当します。

例えば 4 歳の平均身長が 100cm で標準偏差 10cm だったとしますと、身長 80cm の子供は何 SD に位置するでしょうか？

答え－2SD

例えば 18 歳から 74 歳の収縮期血圧平均が 129mmHg で標準偏差が 19.8 であったとします。この集団で血圧の高い方から 2.5%に相当する人は何 SD 以上でしょうか？

正規分布の片端の面積が 2.5%は先に示した通り z=1.96 です。上記公式に当てはめると、1.96＝(X－129)/19.8　X=167.8 mmHg となります。つまりこの集団の上位 2.5%という集団は 168mmHg の血圧をもっており、逆に 97.5%の人達はそれ以下であることを示しています。逆に血圧の低い方はどうでしょうか？1.96=(X + 129)/19.8　X = 90.2 mmHg が－1.96SD 以下、あるいは 2.5%の集団に属すると言えます。あるいは適当に選んだ人が血圧 90 以下である確率は 2.5%とも言い換えられます。



血圧　　90.2　129　167.8

Z　　-1.96SD　1.96SD

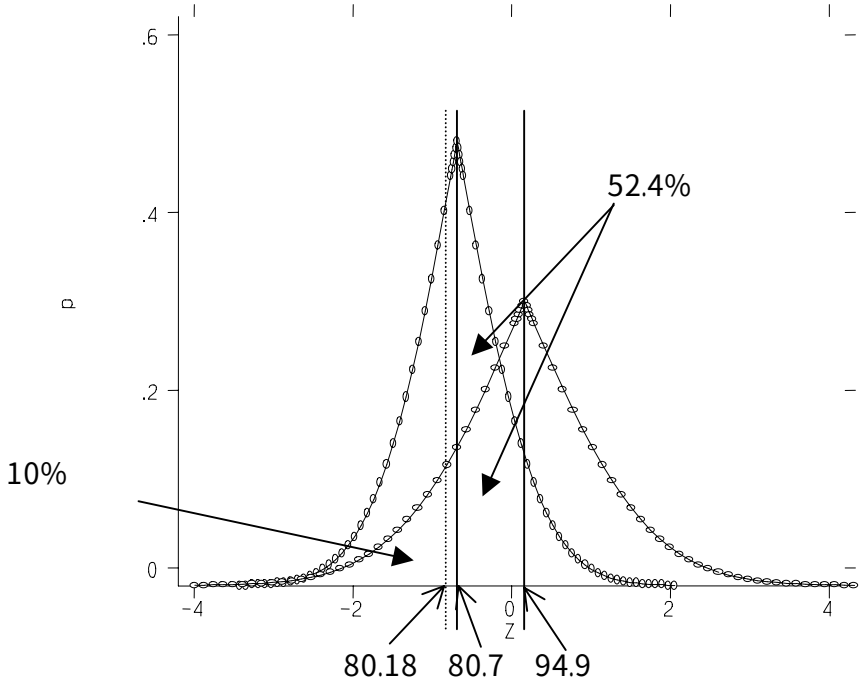
それでは血圧が 150mmHg の人は、この集団内で何%いますか？

Z = 150 – 129 / 19.8　z=1.06,　14.5%

よってこの集団の 14.5%は血圧 150mmHg であると言えます。

ここに正常血圧の集団と高血圧だが抗高血圧薬を服用している集団がいるとします。それぞれの最低血圧平均と標準偏差は、μ_1=80.7, σ_1=9.2, μ_2=94.9, σ_2= 11.5 でした。どのへんをカットオフ値としたらよいでしょうか？

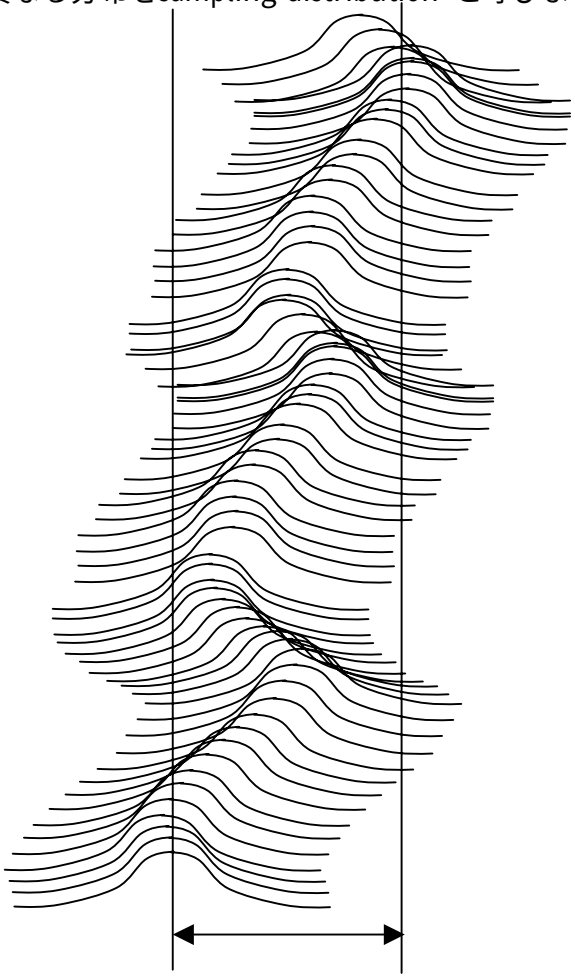
仮に高血圧の集団の左端 10%の集団をとりますと、表より 0.10 に相当する z＝－1.28 です。-1.28 = x – 94.9 / 11.5,　x = 80.18 mmHg,
この血圧は正常群の集団のどこに位置するのでしょうか？z = 80.18 – 80.7 / 9.2 = 0.06,
このｚより左端にある部分の面積は表より 0.476、逆に右にある部分の面積は 0.524 です。この 52.4%は高血圧の下 10%よりも血圧の高い集団であり、偽陽性となります。半分以上が偽陽性となってしまっては、よいカットオフ値とは言えません。もし血圧 90mmHg できったとすると高血圧群の 66.6%が偽陰性となり、33.4%が偽陽性となります。２つの集団がオーバーラップすればする程わけるのが難しくなります。



推論

世界中の人のコレステロールの平均値を知りたいのですが、これは不可能な話です。そこである似た集団でコレステロールの値を測定して代用することにしました。ここで得た平均値$\bar{X}$,は、全体から得られるはずの平均値 μ の Maximum likelihood estimator と呼ばれます。つまり、「多分この 2 つの値は近似しているだろうけど、どの程度近い値かはわからないよ。」ということです。つまり不正確さが含まれる点に注意しなくてはなりません。

とにかく一部をみて全体を予測しようという話ですから、sampling をするときは注意が必要です。例えば対象を 20－70 歳としていても、sampling が 60 歳以上ではコレステロールの値は高くでてしまいます。random かつ十分な数のsamplingが必要となってきます。しかし仮にきちっとした方法で正当にsamplingしても 2 回別々に行なったsamplingに対する平均x_1とx_2は少しずつ異なります。Sampling を何度も繰り返したとして、その少しずつ異なる分布をsampling distribution と呼びます。



上図のように sampling する毎にその平均は少しずつずれることでしょう。しかし大部分が収まるレンジが存在します。それが confidence interval (CI)です。（後述）

20歳から 74 歳男性のコレステロールは平均が 211、標準扁 s は 46 でした。もしこの集団から 25 sample を抽出したとところ平均が 230 以上になる確率はいくつですか？

z = (230 – 211) / (46/ √25) = 2.07　(0.019)

これは 1.9%が 230 を超えていることを示しています。

それでは逆にコレステロールがいくつ以下の時全体の 10%に当たるでしょうか？

10%の面積に相当する z は－1.28 なので

-1.28 = (x – 211) / 9.2　　x = 199.2

となります。よって 25 sample のうち 10%は 199.2 以下の平均 n を持つと言えます。

それでは同様に 25 sample を抽出した際、この集団の 95%はどの範囲に収まるでしょうか？

95%ということは両端が 2.5%ずつであり、これは z＝1.96 となります。つまり

-1.96 ≦z≦1.96
-1.96 ≦x – 211 / 9.2 ≦1.96
193.0 ≦x≦229.0

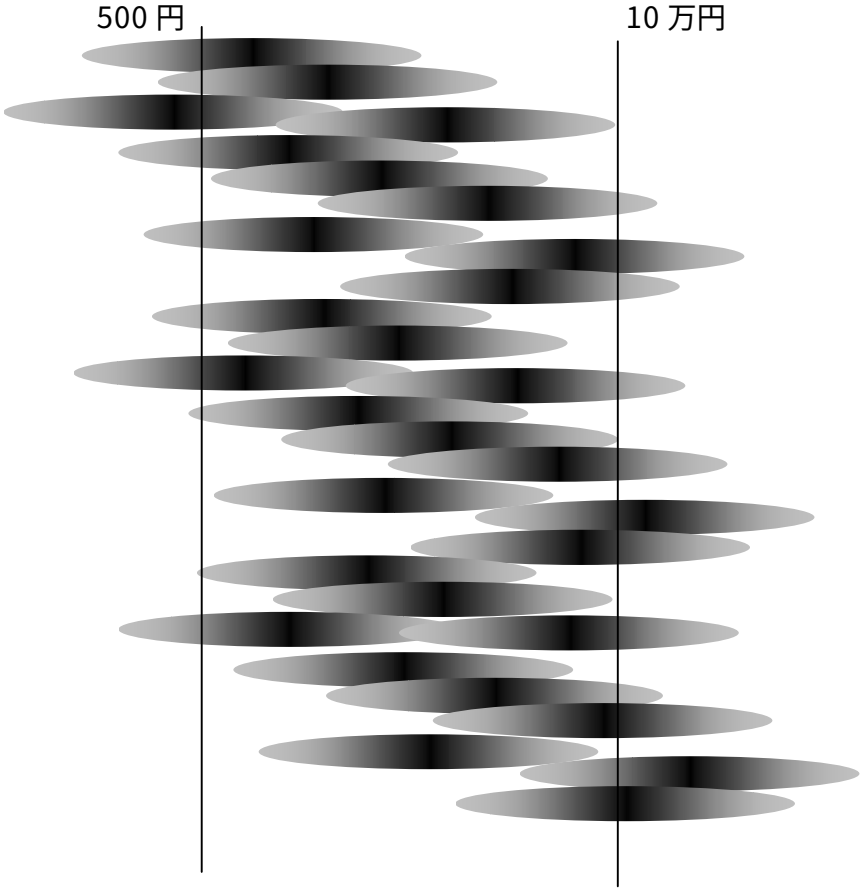
つまり 25 の標本を選んだときその平均は 95%の確率で 193.0 から 229.0 の間にくると言えます。ですから 25 人を無作為に選んだとき、193.0 から 229.0 の間に来なかった場合、起こりにくいこととして(rare events)標本の抽出方法に何か問題があったのではないかと考えるべきです。もし 25 ではなくて 10 を選んだとき、同様に計算して 182.5 ≦x≦239.5 となり、その幅は広がり, もし 100 の標本を選んだときには 202.0 ≦x≦220.0 となりその幅は狭まります。つまり標本抽出によるばらつきが少なくなるということです。標本抽出を多くすればする程、真の 95%の幅は狭くなり、無限大にすることによって真の平均値に集約することになります。

それでは逆に何人の人を対象に調査すればその 95%が平均の±5 に収まるでしょうか？

同様に 95%ということは-1.96 ≦z≦1.96 ということであり、z = X - μ / σ /√n ですから、-1.96 ≦X - μ / σ /√n≦1.96 ということになります。さて P(μ - 5 ≦X≦μ + 5) = 0.95 なので、P(- 5 ≦X - μ≦+ 5) = 0.95 となりますから、更に P(- 5 / σ /√n ≦ X - μ/ σ /√n ≦ + 5/ σ /√n) = 0.95 となって、σ ＝46 ですから、1.96= 5/ 46 /√n = 5√n / 46, 1.96 x 46 / 5 =√n, n = 325.2.
つまり 326 の標本を選らんでくれば、95%の確率で、標本の平均は 206 – 217 の範囲に収まります。

信頼区間(confidence interval)

今までは母集団の平均と分散がわかっている状態について勉強してきましたが、今度はわからない場合です。日本人の「財布の中にいくらの現金持っているか」を調べるとします。日本人全員について調べることはできません。よって本当の母集団については永遠に解りません。推論するしかないのです。そこである街角でアンケートをすることにしました。10 人に聞くよりも、100 人、1000 人と数が多い程、母集団の値に近くなるでしょう。また sampling も場所を変えて行なうべきかもしれません。たとえば銀座の 3 丁目の交叉点でのアンケートと北海道宗谷岬でのアンケートの結果は違うでしょう。そこで日本全国 100 ヶ所を選び、それぞれで 1000 人ずつより解答を得ようと思います。中には極端に安いところがあるでしょうし、極端に高いところもあるでしょう。両極端な 2.5%ずつを除いた残り 95%が収まる範囲をもって母集団の近似すると考えます。すなわち 100 ヶ所のうち 95 ヶ所の収まる範囲に母集団の真の平均が収まるであろうと推論します。



例えば極端な平均を除いた 95 ヶ所での平均は 500 円から 10 万円の間に収まったとします。この範囲内に 95%の確率で日本人全体の財布の中の現金の値段が収まると考えます。誰も知らない値ですから、このように一定の範囲を設けて推論するしかないので

す。

　少し難しく解説すると、信頼区間とは「ある標本をとった場合、その平均が母集団の平均とどれくらい隔たっていて、その隔たりを示す数値がどれくらい信用できるか」を示す事にもなります。

$Z = (X - \mu)/ \sigma /\sqrt{n}$

は n が十分大きければ standard normal random variable です。正規分布では観察した標本の 95%は－1.96 から 1.96 の間に存在します。言いかえれば

$P(-1.96 \leqq z \leqq 1.96) = 0.95$
ｚを $X - \mu / \sigma /\sqrt{n}$ と置き換えて
$P(-1.96 \leqq X - \mu / \sigma /\sqrt{n} \leqq 1.96) = 0.95$
少し変えて
$P(-1.96 \sigma /\sqrt{n} \leqq X - \mu \leqq 1.96 \sigma /\sqrt{n}) = 0.95$
更に少し変えて
$P(X - 1.96 \sigma /\sqrt{n} \leqq \mu \leqq X + 1.96 \sigma /\sqrt{n}) = 0.95$

結論として平均値μは$(X - 1.96 \sigma /\sqrt{n}$ 、$X + 1.96 \sigma /\sqrt{n})$の間に 95%の確率で存在すると推論できます。
　例えばコレステロールの分布を考えてみた場合、平均値μに対して標準偏差が 46 であったとしますと、標準を抽出する前に 95%の信頼（確率とは言わない）で平均μは下記の範囲に存在することになります。

$(X - 1.96 \times 46 /\sqrt{n}、X + 1.96 \times 46 /\sqrt{n})$

もしもこの集団より 12 人を抽出したところ、平均値が 217 であったとしますと、
$(217 - 1.96 \times 46 /\sqrt{12}、217 + 1.96 \times 46 /\sqrt{12})$
(196, 243)
「95%の信頼をもって平均値μは 196 と 243 の間に存在する」と推論できます。

One-Sided Confidence Intervals

ある状況ではある集団の極端な値というよりは、異常に高い、あるいは異常に低いもしばしば問題になります。高濃度の鉛に暴露された 6 歳以下の子供のヘモグロビン分布を検討してみました。この分布は平均値μと標準偏差が 0.85 です。鉛によってヘモグロビンレベルが低下することがわかっていますから、上の値がわかれば十分です。

95%から外れる、言いかえれば高い値 5%がヘモグロビンのいくつに当たるかを調べればよりわけです。

$P(-1.645 \leqq Z) = 0.95$
$P(-1.645 \leqq X - \mu / \sigma / \sqrt{n}) = 0.95$
$P(\mu \leqq 1.645 + X + \sqrt{n} / 0.85) = 0.95$

もし鉛に暴露された子供 75 人についてヘモグロビンを測定したところ、平均値が 10.6 だったとします。上記式にあてはめて、
$\mu \leqq 1.645 + 10.6 + \sqrt{75} / 0.85$
$\mu \leqq 10.8$

95%の信頼をもって本当のヘモグロビン平均値はせいぜい 10.8 であると結論できます。We are 95% confident that the true mean hemoglobin level for this population of children is at most 10.8%/dl.

Student's t Distribution

平均値が解っていないときでも標準偏差が解っていれば何とかなりました。でも実際平均値が解らないときは標準偏差も解らないのが普通です。この場合でも信頼区間は従来行ってきたように計算します。しかし今度は正規分布の分散を用いずに、student's t distribution を用います。

最初は同様に$Z = (X - \mu) / \sigma / \sqrt{n}$ を用います。この場合、σではなく、理論代用(logical to substitute) s を標準偏差として用います。$t = (X - \mu) / (s / \sqrt{n})$ となります。しかしこの比はstandard normal distribution を示しません。また n が小さいときは s もσのように信頼できるものではありません。ある集団から適当に抽出した標本 n に対してランダム変数の分布はn-1 の自由度をもつStudent's t distribution として知られています。これをt_{n-1} と表します。正規分布と同様にt 分布は左右対称であり、カーブの下の総和は 1 になります。しかしながら正規分布に比べると幾分つぶれており端っこ

は厚いのが特徴です。

自由度が異なるとｔ分布も異なってきます。自由度の小さいものは横に広がり、自由度が大きくなると正規分布に近付きます。測定値が多くなるとｓはσに近付くからです。

自由度が異なるとt分布も異なる為非常に大きなテーブルとなってしまうので、通常代表的な数値しか示してありません。例えば自由度が 10 の時t_{n-1}は 2.228 で、これより外側にある部分の面積は 2.5%となります。正規分布ではz = 1.96 が端っこの 2.5%に一致しました。よってt分布でもｎが増えると正規分布に近似します。実際ｎが 30 以上になるとｔ分布を正規分布で代用できます。例えばアルミニウムを含む制酸剤を投与された乳児 10 人について考えてみましょう。この制酸剤を投与された乳児全員については何もわかりませんが、調査した 10 人のアルミニウムの値の平均 37.2 と標準偏差 7.13 はすぐに計算ができます。ｔ分布において、10 人を対象としており、その自由度は 9 であり、95%信頼区間は－2.262 から 2.262 の間に位置します。よってこの母集団の平均値μに対する信頼区間は

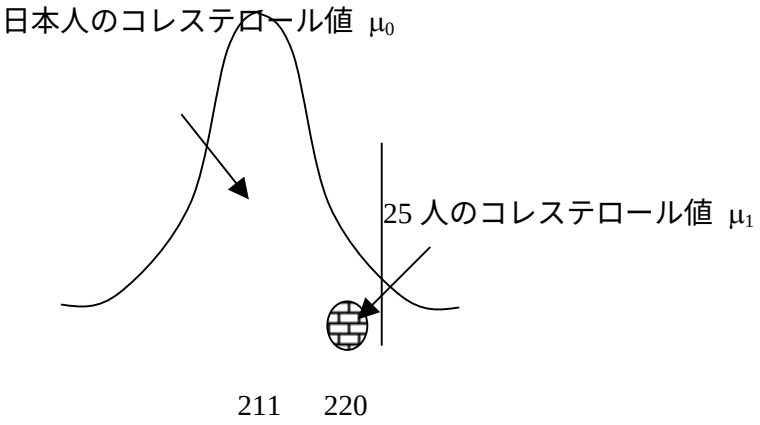
$(x - 2.262\ s / \sqrt{n}\ ,\ x + 2.262\ s / \sqrt{n})$で表され、これに実数を当てはめると
$(37.2 - 2.262 \times 7.13 / \sqrt{10}\ ,\ 37.2 + 2.262 \times 7.13 / \sqrt{10})$、
(32.1, 42.3)

となり、アルミニウムを含む制酸剤を投与された乳児の血中アルミニウム値の平均値は 95%の信頼度をもって 32.1 から 42.3 の間に存在します。もしもこれが正規分布であったとして計算すると（z=1.96）、32.8 から 41.6 の間に平均が存在することになります。10 人からのデータでも正規分布の値に近似するのがわかります。先にも述べたとおり、同じ信頼度であればｎが大きくなればなるほど信頼区間は狭くなります。

仮説の検証

もしも日本人のコレステロールの平均値が 211 mg/dl であったとします。心筋梗塞を発症した患者さん 25 人のコレステロールの平均は 220 mg/dl であったとします。これは高いのでしょうか？それとも正常範囲内なのでしょうか？

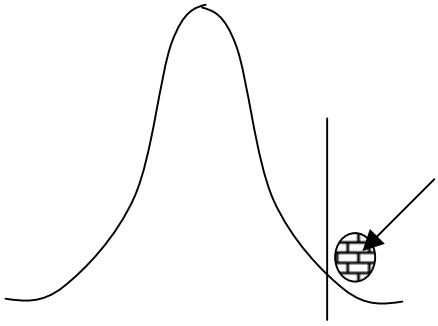
まずは正常範囲内であると仮定します。つまり「25 人の平均は母集団の平均と一致する」という仮説です。



Null hypothesis ：H_0: $\mu_0 = \mu_1$

あるラインを決め、それより内側であれば 25 人のコレステロール値は日本人のそれと同じであると考えます。一般的には片側 5%あるいは両側 5%を適応します。それぞれに対応する z は 1.645 あるいは 1.96 です。つまり日本人の平均＋1.96SDを 25 人の平均が超えていればHoを棄却して 2 つの平均は異なる､もしも超えていなければH_0 をaccept して２つの平均は同じ、つまり 25 人のコレステロール値は正常範囲内であると結論します。この$p < 0.05$ という基準はあくまで統計学的合意であって不変の真理を表すものではありません。よって臨床的意義と一致するとは限らないのです。特にsample 数が少なかったとき、p値による統計学的有意性を後ろ盾するパワーが弱いため、「sample 数が少ない為十分な結論を述べる事はできない」という奇妙な結論になってしまいます。Type I error, type II error, power, sample size に関しては臨床試験の項を参照してください。

一方 25 人の平均は母集団の範囲を逸脱するとも考えられます。Alternative hypothesis



$H_A: \mu_0 \neq \mu_1$
となります。

そしてこれからどちらの仮説が正しいか検討することになります。
　もし比較する集団の標準偏差が分からない場合には測定した値より標準偏差を割り出しこれをもって代用します。ここでは２つの集団が同じと仮定しているので、標準偏差も等しいことになります。もし母集団が正規分布を示すのであれば、

$t = (X - \mu_0) / s / \sqrt{n}$

で表され t 分布は自由度 n-1 となります。この場合表をもって比較します。

$H_0: \mu = 211$ mg/dl
$H_A: \mu \neq 211$mg/dl

かどうかを証明します。
25 人調査したところ平均が 220 で SD が 46 だったとします。
$z = (X - \mu_0) / \sigma / \sqrt{n}$
$z = (220 - 211) / 46 / \sqrt{25}$
$= 0.98$
この値は上で示したライン、1.645 あるいは 1.96 より小さく統計学的に有意ではありません。つまり、この測定した集団のコレステロール平均が母集団のそれと異なると結論するには十分な証拠がないということになります。これは裁判に似ています。つまりいろいろな証拠をあつめてやりあうのですが、その本当の真偽は分からないまま裁判官は決断を下すこともあります。つまり、本当は殺人犯なのに証拠不足で無罪となってしまうこともあるわけです。
　それでは 25 人の平均がいくつ以上だったら有意差がでると思いますか？

1.96 =(x - 211) / 46 /√25
x＝229 mg/dl となります。

One-Sided Tests of Hypothesis

仮説を検証する前に、平均μ_0からどれくらい隔たっているかをみる際、両側で検討するか片側で検討するかを決めておかなくてはなりません。この決定はサンプルを抽出する前に行われなくてはなりません。サンプルを集めた結果で決めるのはうアンフェアだからです。例えば鉛に暴露された子供のヘモグロビンの値が低くなることはあっても高くなることはないので、このような場合は片側だけ検討すれば十分ということになります。

H_0: $\mu \leqq \mu_0$
H_A: $\mu > \mu_0$

鉛に暴露された子供のヘモグロビンに話を戻しましょう。もし 6 歳の子供のヘモグロビンの平均が 12.29 g/dl, σ = 0.85 g/dl であったとしますと、期待する値はこれより低くので、

H_0: $\mu \geqq$12.29 g/dl
H_A: $\mu <$ 12.29 g/dl

今回は片側の検討なのでα = 0.05 のところをとります。σが分かっているので正規分布の表を用いればよいことになります。適当に抽出された鉛に暴露された 74 人の子供のヘモグロビンのレベルは 10.6 g/dl でした。

z =X - μ_0 / σ /√n
z =10.6 - 12.29 / 0.85 /√74
= -17.10 <0.001

よって我々は
H_0: $\mu \geqq$12.29 g/dl
という仮設を否定するこのになります。一般に z が－1.645 より小さければ仮説を否定できるということです。

One sided の場合片側 5%以上のところに位置すると有意差ありとなるのですが、Two sided の場合は片側 2.5%となるので、One sided で有意であっても、Two sided

で有意でない場合も当然ありえることになります。そのため、One sided test をきらう編集者がいます。

慢性膵炎の患者さんの血糖値を測定し以下の値を得ました。血糖の正常値が 100mg/dl としたとき、この集団の血糖値は正常範囲でしょうか？
117, 119, 99, 114, 120, 104, 88, 114, 124, 116, 101, 121, 152, 90, 125, 114, 95, 117

まずはデータを入力します。

```
. list

        BS
 1.    117
 2.    119
 3.     99
 4.    114
 5.    120
 6.    104
 7.     88
 8.    114
 9.    124
10.    116
11.    101
12.    121
13.    152
14.     90
15.    125
16.    114
17.     95
18.    117
```

上のデータを示す集団が variance の不明な既知の値と異なるかどうかを検定します。

```
. ttest  BS=100

One-sample t test

------------------------------------------------------------------------------
```

```
Variable |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
      BS |      18    112.7778    3.559536    15.10183    105.2678    120.2877
------------------------------------------------------------------------------
Degrees of freedom: 17

                     Ho: mean(BS) = 100

   Ha: mean < 100           Ha: mean ~= 100          Ha: mean > 100
      t =   3.5897              t =   3.5897            t =   3.5897
  P < t =   0.9989        P > |t| =   0.0023        P > t =   0.0011
```

「18 人の血糖の平均μ_1 = 100 mg/dl である」がH_0だったわけですが、上の中央では「これを否定することができる」とでています。つまり「18 人の血糖は正常値100mg/dlとp=0.0023 の統計学的有意差をもって異なる」が結論です。上の血糖値をみると 100 より高い値だけでなく低い値もあるのでtwo sided t-test を行なうのが妥当でしょう。

先の血糖値の問題で、95%信頼区間は 105.2678 ―120.2877 と STATA は示しています。つまり慢性膵炎の異なる 100 の集団に対して血糖値を調べたとき 95 回の平均値はこの範囲内に収まる、あるいは慢性膵炎の真の平均値は 95%の確率でこの範囲にあるといえます。

row data を持ち合わせていない場合でも検討することができます。観察数、平均、SD、母集団（比較しようとする）の平均値の順で直接データを入力します。

```
. ttesti 18 112.8 15.1 100

One-sample t test

------------------------------------------------------------------------------
       |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       x |      18       112.8    3.559104        15.1    105.2909    120.3091
------------------------------------------------------------------------------
Degrees of freedom: 17
```

```
                    Ho: mean(x) = 100

   Ha: mean < 100          Ha: mean ~= 100          Ha: mean > 100
      t =   3.5964             t =   3.5964             t =   3.5964
  P < t =   0.9989         P > |t| =   0.0022         P > t =   0.0011

.
```

2 つの集団の比較

今までは 1 つの集団の平均を母集団の平均μ_0 と比較してきました。しかし現実問題として２つの異なる集団を比較することの方が多いのではないでしょうか。でも２つの集団を比較する方法は１つの場合と似ています。まず２つ集団が等しいという仮説から始めます。次にsignificance p のレベルを決定してone-sided かtwo-sided に決定します。そして２つがpaired or unpaired を決定します。

Paired Samples

最もよくある Paired Samples の例はある処理を行う前と行った後の比較です。もう１つのパターンは１つの集団がもう１つの集団と性や年齢が合っているかどうかみる場合です。つまり Paired Samples の利点は比較をより正確に行うことができることにあります。つまり少ない sample 数でも小さな差でも統計学的有意性を証明しやすくなる点です。

One sided test	Two sided test
$H_0 : \mu_1 - \mu_2 \geqq 0$	$H_0 : \mu_1 - \mu_2 = 0$
$H_A : \mu_1 - \mu_2 < 0$	$H_A : \mu_1 - \mu_2 \neq 0$

それぞれのペアの差は

$d_n = x_{n1} - x_{n2}$で表されるとします。

各患者の差の平均をdとすると、SDは平均dからそれぞれの測定差d_iを引いたものの 2 乗をn-1 で割って平方根にしたものです。もし本当の２つの値の違いを

$\delta = \mu_1 - \mu_2$としますと

$H_0 : \delta = 0$

と書き換えられます。H_0 は下記の式で検証できます。

$t = (d - \delta) / (SD / \sqrt{n})$

$SD/\sqrt{n}$は d のstandard error でしたよね。もしこの集団が正規分布をしてH_0が本当であるとしたら、n-1 の自由度を示す分布になります。Tの値を表より換算してpの値を得ます。この値が・より小さければH_0 を否定できますし、この値がδより大きければH_0を否定できません。

STATAによるpaired t test

手で計算すると非常に複雑に感じられますが、コンピュータを用いると t test は非常に簡単です。たとえば 20 人の喘息患者さんにβ刺激剤吸入約を使用し 1 秒率の変化をみたとします。

患者 ID	吸入前	吸入後
1	88	92
2	91	90
3	75	92
4	63	91
5	68	67
6	60	68
7	69	64
8	72	84
9	70	84
10	69	74
11	72	89
12	78	95
13	71	78
14	73	73
15	67	70
16	71	80
17	68	72
18	70	75
19	64	70
20	72	89

上のデータを用いこの新しい吸入薬が喘息患者さんの 1 秒率を有意に改善したかどうか two sided paired t test で検討してください。

まずは STATA にデータを入力します。

```
. list
```

```
        pre     post
  1.     88       92
  2.     91       90
  3.     75       92
  4.     63       91
  5.     68       67
  6.     60       68
```

```
 7.      69       64
 8.      72       84
 9.      70       84
10.      69       74
11.      72       89
12.      78       95
13.      71       78
14.      73       73
15.      67       70
16.      71       80
17.      68       72
18.      70       75
19.      64       70
20.      72       89
```

そして以下のようにコマンドします。

```
. ttest pre=post

Paired t test

------------------------------------------------------------------------------
Variable |     Obs       Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
     pre |      20      71.55    1.648724    7.373316    68.09918    75.00082
    post |      20      79.85    2.221042    9.932801    75.20131    84.49869
---------+--------------------------------------------------------------------
    diff |      20       -8.3    1.813836     8.11172    -12.0964   -4.503598
------------------------------------------------------------------------------

                Ho: mean(pre - post) = mean(diff) = 0

  Ha: mean(diff) < 0        Ha: mean(diff) ~= 0       Ha: mean(diff) > 0
      t =  -4.5759              t =  -4.5759              t =  -4.5759
   P < t =   0.0001         P > |t| =   0.0002         P > t =   0.9999
```

Two sided t test は上の中央に相当しますが、p=0.0002 であり統計学的な有意性が証明されました。この新しい吸入薬は喘息患者さんの 1 秒率を明らかに改善します。

上のデータが 12 人の吸入前後の変化でなく、24 人の喘息患者さんに対してであって 12 人には治療せず、12 人には吸入薬を使用したものだとするとどうでしょうか？

この場合 paired test でなく、unpaired test となります。

```
. ttest pre=post, unpaired

Two-sample t test with equal variances

------------------------------------------------------------------------------
Variable |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
     pre |      20       71.55    1.648724    7.373316    68.09918    75.00082
    post |      20       79.85    2.221042    9.932801    75.20131    84.49869
---------+--------------------------------------------------------------------
combined |      40        75.7    1.518349    9.602884    72.62885    78.77115
---------+--------------------------------------------------------------------
    diff |              -8.3    2.766101             -13.89968   -2.700321
------------------------------------------------------------------------------
Degrees of freedom: 38

                    Ho: mean(pre) - mean(post) = diff = 0

     Ha: diff < 0                 Ha: diff ~= 0                 Ha: diff > 0
       t =  -3.0006                 t =  -3.0006                 t =  -3.0006
   P < t =   0.0024         P > |t| =   0.0047         P > t =   0.9976
```

有意ではあります。しかし paired t test よりｚは小さくなってしまいました。同じ人の変化をみる方が powerful です。

Independent Samples

健康な子供とcystic fibrosis (CF) の子供の血清鉄の値を比較してみましょう。２つの集団は互いに独立していて正規分布を示すとします。ここでCF の子供の血清鉄値の平均はμ_1 とし、健康小児のそれをμ_2とします。で、この２つの集団が同じであるという仮説から始めます。

$H_0 : \mu_1 - \mu_2 = 0$

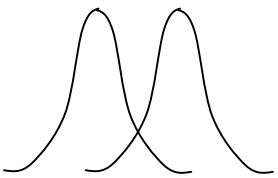
あるいは

$H_0 : \mu_1 = \mu_2$

$H_A : \mu_1 \neq \mu_2$ (alternative hypothesis)

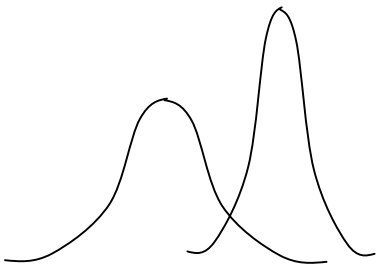
		Group 1	Group 2
Population	Mean	μ_1	μ_2
	Standard deviation	σ_1	σ_2
Sample	Mean	x_1	x_2
	Standard deviation	s_1	s_2
	Sample size	n_1	n_2

独立したサンプルを比較するとき、２つの異なるシナリオがあります。最初にもとの集団の variance が分かっていて同じである事(equal variances) が予測されるとき、two-sample t test をすればよりわけです。一方 variance が同じでないこと(unequal variances)が予測されるとき、two-sample t test は使えません。



Equal variances

Two sample t test

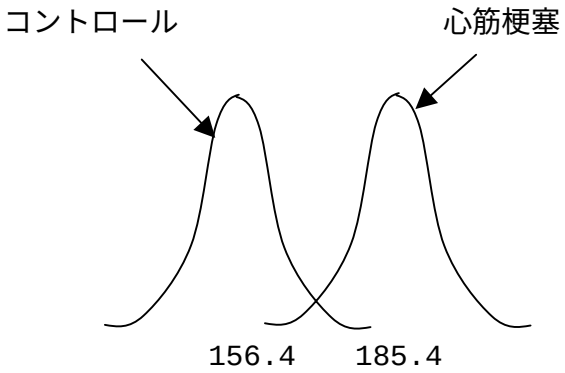


unequal variances

下は 20 人の心筋梗塞を発症した患者さんの血清コレステロール値とコントロールの人のそれを示しています。Variance が等しいとして心筋梗塞患者さんで血清コレステロールがコントロールと異なるか検討してください。

```
. list

        AMI     Chole
  1.      0       156
  2.      0       157
  3.      0       183
  4.      0       130
  5.      0       129
  6.      0       133
  7.      0       182
  8.      0       175
  9.      0       199
 10.      0       134
 11.      0       165
 12.      0       142
 13.      0       120
 14.      0       183
 15.      0       145
 16.      0       173
 17.      0       172
 18.      0       155
 19.      0       173
 20.      0       122
 21.      1       176
 22.      1       187
 23.      1       190
 24.      1       188
 25.      1       172
 26.      1       161
 27.      1       122
 28.      1       103
 29.      1       154
```

```
 30.       1      138
 31.       1      167
 32.       1      189
 33.       1      177
 34.       1      169
 35.       1      203
 36.       1      122
 37.       1      240
 38.       1      283
 39.       1      299
 40.       1      268

. ttest Chole, by(AMI)

Two-sample t test with equal variances

------------------------------------------------------------------------------
   Group |    Obs       Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
      0 |     20      156.4    5.242739    23.44624    145.4268    167.3732
      1 |     20      185.4    11.71939    52.41073     160.871     209.929
---------+--------------------------------------------------------------------
combined |     40      170.9    6.748485    42.68117    157.2499    184.5501
---------+--------------------------------------------------------------------
    diff |              -29    12.83863             -54.99046   -3.009544
------------------------------------------------------------------------------
Degrees of freedom: 38

                 Ho: mean(0) - mean(1) = diff = 0

     Ha: diff < 0                 Ha: diff ~= 0                 Ha: diff > 0
       t =  -2.2588                 t =  -2.2588                 t =  -2.2588
   P < t =   0.0149             P > |t| =   0.0297             P > t =   0.9851
```



心筋梗塞を発症した患者さんの血清コレステロール値は測定した健康人のそれとの間に有意差を認めました。つまり心筋梗塞を発症した患者さんの血清コレステロール値は正常範囲を逸脱していたと結論できます。

Equal variance の項で STATA を用いて解析した心筋梗塞と血清コレステロールの例題を unequal variance として解析しなおしてください。

```
. ttest Chole, by(AMI) unequal welch

Two-sample t test with unequal variances

------------------------------------------------------------------------------
   Group |     Obs       Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       0 |      20      156.4    5.242739    23.44624    145.4268    167.3732
       1 |      20      185.4    11.71939    52.41073     160.871     209.929
---------+--------------------------------------------------------------------
combined |      40      170.9    6.748485    42.68117    157.2499    184.5501
---------+--------------------------------------------------------------------
    diff |              -29    12.83863             -55.33898   -2.661015
------------------------------------------------------------------------------
Welch's degrees of freedom:  27.0817

                  Ho: mean(0) - mean(1) = diff = 0

     Ha: diff < 0                 Ha: diff ~= 0                 Ha: diff > 0
       t =  -2.2588                 t =  -2.2588                 t =  -2.2588
   P < t =   0.0161         P > |t| =   0.0322         P > t =   0.9839
```

心筋梗塞患者さんでは有意に血清コレステロールが高い傾向にありました。equal variance として計算したとこと同じ結果になりましたが、p値は大きくなってしまいました。

Nonparametric Methods

今までは、正規分布かそれに近いものでした。このようなものを検証する方法は parametric と呼ばれます。一方正規分布を示さない場合には nonparametric methods を使用します。

The sign test

The sign testはpaired t test でdistribution が正規分布を示さないときに使用します。サンプル数nが小さければ、Z_+は正規分布になりません。もしもnが 10 であったとすると、正規分布になることのほうが希でしょう。

例えば 10 人の喘息患者さんに気管支拡張剤を吸入してもらい、1 秒率がどの程度改善するかを検討するとします。Paired t test と sign test を行なって比較してみてください。

```
. gen d=post - pre

. list

        pre     post        d
  1.     88       90        2
  2.     95       97        2
  3.     90       95        5
  4.     76       82        6
  5.     65       72        7
  6.     78       86        8
  7.     82       73       -9
  8.     79       90       11
  9.     75       88       13
 10.     63       99       36
```

単純に１秒率が良くなったら＋、悪くなったらーで表します。Null hypothesis では＋とーの数が一致するはずです。つまり＋になる確率は 1/2 であり、±で示されるものの分布は binomial distribution を示すので、平均は np = n/2, variance は√np(1-p) = √n/4 となります。ですから観察された＋の割合が n/2 と同じかどうかを検討すればよいことになります。

Z = [+の数 – (n/2)]/√(n/4)

上で 9 人は１秒率を示す数値が多かれ少なかれ増えているので＋となります。よって

Z = [9 – 5]/√(10/4) = 2.53

Z は 1.96 より大きいので null hypothesis を棄却して「吸入薬により有意に改善した」と結論します。STATA で計算するとどうなるでしょうか？

```
. signtest pre=post

Sign test

    sign |   observed   expected
---------+------------------------
positive |         1         5
negative |         9         5
    zero |         0         0
---------+------------------------
     all |        10        10

One-sided tests:
Ho: median of pre - post = 0 vs. Ha: median of pre - post > 0
    Pr(#positive >= 1)
    = Binomial(n = 10, x >= 1, p = 0.5) =  0.9990

Ho: median of pre - post = 0 vs. Ha: median of pre - post < 0
    Pr(#negative >= 9)
    = Binomial(n = 10, x >= 9, p = 0.5) =  0.0107

Two-sided test:
Ho: median of pre - post = 0 vs. Ha: median of pre - post ~= 0
    Pr(#positive >= 9 or #negative >= 9)
    = min(1, 2*Binomial(n = 10, x >= 9, p = 0.5)) =  0.0215
```

two sided test で有意差ありです。

```
. ttest pre=post

Paired t test
```

```
------------------------------------------------------------------------------
Variable |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
     pre |      10        79.1    3.240199    10.24641    71.77016    86.42984
    post |      10        87.2    2.931818    9.271222    80.56777    93.83223
---------+--------------------------------------------------------------------
    diff |      10        -8.1    3.640665    11.51279   -16.33576    .1357573
------------------------------------------------------------------------------

                 Ho: mean(pre - post) = mean(diff) = 0

  Ha: mean(diff) < 0         Ha: mean(diff) ~= 0         Ha: mean(diff) > 0
      t =  -2.2249               t =  -2.2249               t =  -2.2249
   P < t =   0.0266          P > |t| =   0.0531          P > t =   0.9734
```

通常の paired t test では two sided test において有意差を検出できませんでした。

The Wilcoxon Signed-Rank test

Sign test だとどんな分布を想定してもできますが、どれくらい違うかという情報を無視しています。そういったわけで、Sign test はあまり使用されません。そこで Wilcoxon Signed-Rank test は独立した 2 つの集団からのサンプルを比較するのに使用されます。Sign test のように個々の集団を個別に検証するのではなく、個々の観察のペアとして検証します。更に Wilcoxon Signed-Rank test では差を定量化しもします。

例えばある疾患に対する治療効果をみるためにプラシーボと薬物治療群にわけます。そしてその差をとります。符号（ﾌﾟﾗｽﾏｲﾅｽ）は無視して、絶対値で比較して小さい方から順番をつけます。更にそれぞれに新たに 1 から順番に番号をつけてやります。もし同点が 2 個あればその平均をとります。更にもとにあった符合を付け直して、signed rank とします。そして＋のものの合計と－のものの合計を出します。また符号（ﾌﾟﾗｽﾏｲﾅｽ）は無視して小さい方の合計を T とします。Null hypothesis では 2 つの集団は等しい中央値を持つと仮定しているので、プラスの合計とマイナスの合計の絶対値がほぼ一致するかどうか検討します。

$z_T = (T \cdot \mu_T) / \sigma_T$
$\mu_T = n(n+1) / 4$

$\sigma_T = \sqrt{n(n+1)(2n+1) / 24}$

先の喘息の例題で､同じデータを 2 群に分けた合計 20 人の喘息患者さんと想定します。Pre はプラシーボを与え、post はβ刺激薬を与えたものとします。Signrank test を行なってβ刺激薬の効果についてコメントしてください。

```
. signrank pre=post

Wilcoxon signed-rank test

   sign |     obs   sum ranks    expected
---------+---------------------------------
positive |       1          7       27.5
negative |       9         48       27.5
   zero |       0          0          0
---------+---------------------------------
    all |      10         55         55

unadjusted variance      96.25
adjustment for ties      -0.12
adjustment for zeros      0.00
                  ----------
adjusted variance        96.12

Ho: pre = post
           z =  -2.091
   Prob > |z| =   0.0365
```

この吸入薬は有意に喘息患者さんの１秒率を改善したと結論できます。

The Wilcoxon Signed-Rank sum test (Mann-Whitney test)

The Wilcoxon Signed-Rank sum test (Mann-Whitney test)は非独立集団に対して適応されます。よって unpaired t test の nonparametric 板といったところで、正規分布する必要もなく、標準偏差も一致する必要がありません。しかしながらその分散は同じ形として扱います。

```
. ranksum EFV, by(drug)

Two-sample Wilcoxon rank-sum (Mann-Whitney) test

   drug |     obs    rank sum    expected
---------+---------------------------------
      0 |      10       83.5        105
      1 |      10      126.5        105
---------+---------------------------------
combined |      20        210        210

unadjusted variance     175.00
adjustment for ties      -0.92
                  ----------
adjusted variance       174.08

Ho: EFV(drug==0) = EFV(drug==1)
           z =  -1.630
   Prob > |z| =   0.1032
```

この検定では Wilcoxon signed-rank test で有意であったものを有意でないと判定しています。いずれにしても paired test より unpaired test は power の面で劣ります。