Lectures Notes: Review*(STA4210)

Rohit Patra

August 31, 2018

1 Expected value

The distribution of a random variable X contains all of the probabilistic information about X. The entire distribution of X, however, is usually too cumbersome for presenting this information.

Thus we need some *summary* measures, like the center of the distribution, spread of the distribution.

Intuition: The *expected value* of a random variable indicates its (weighted) average.

Example: How many heads would you expect if you flipped a coin twice?

Let X = number of heads. Then $X = \{0, 1, 2\}$. Here $p(0) = \frac{1}{4}, p(1) = \frac{1}{2}, p(2) = \frac{1}{4}$. [Draw p.m.f!]

Weighted average $= 0 \cdot \frac{1}{4} + 1 \cdot \frac{1}{2} + 2 \cdot \frac{1}{4} = 1$.

Definition: Let X be a *bounded* random variable assuming the values $x_1, x_2, x_3, \ldots$ with corresponding probabilities $p(x_1), p(x_2), p(x_3), \ldots$. The *mean* or *expected value* of X is defined by

$$\mathbb{E}(X) = \sum_k x_k \cdot p(x_k).$$

Interpretations:

(i) The expected value measures the center of the probability distribution – center of mass.

---

*Notes adapted and borrowed from classes taught by Deb Burr and Larry Winner at University of Florida

(ii) Long term frequency.

Expectations can be used to describe the potential gains and losses from games.

Example: Roll a die. If the side that comes up is odd, you win the USD equivalent of that side. If it is even, you lose USD 4.

Solution: Let X = your earnings. Thus,

Table 1: Distribution of r.v X

X=1	$\mathbb{P}(X = 1) = \mathbb{P}(\{1\}) = 1/6$
X=3	$\mathbb{P}(X = 3) = \mathbb{P}(\{3\}) = 1/6$
X=5	$\mathbb{P}(X = 5) = \mathbb{P}(\{5\}) = 1/6$
X=-4	$\mathbb{P}(X = -4) = \mathbb{P}(\{2, 4, 6\}) = 3/6$

$$\mathbb{E}(X) = 1 \cdot \frac{1}{6} + 3 \cdot \frac{1}{6} + 5 \cdot 16 + (-4) \cdot \frac{1}{2} = \frac{1}{6} + \frac{3}{6} + \frac{5}{6} - 2 = -\frac{1}{2}.$$

Definition: Let X be a discrete random variable whose p.m.f is p. Then the *mean, expectation, expected value* of X is defined to be

$$\mathbb{E}(X) = \sum_x x \cdot \mathbb{P}(X = x). \tag{1}$$

Note that the expectation of a random variable X depends **only** on the distribution of X.

1.1 Expectation of a function of a random variable

Let X be a r.v assuming values $x_1, x_2, \ldots$ with corresponding probabilities $p(x_1), p(x_2), \ldots$.

For any function g, what is $\mathbb{E}[g(X)]$?

The mean or expected value of $g(X)$ can be found by applying the definition of expectation to the distribution of $g(X)$, i.e., let $Y = g(X)$, determine the probability distribution of Y, and then determine $\mathbb{E}(Y)$ by applying the definition (e.g., (1)).

Example: Roll a fair die. Let X = # of dots on the side that comes up. Let

$$Y = \begin{cases} 1 & \text{if } X \text{ is odd} \\ 0 & \text{if } X \text{ is even.} \end{cases}$$

Calculate $\mathbb{E}(Y)$.

Solution: Y takes two values 0,1 with probability 1/2 each. Thus, $\mathbb{E}(Y) = 1/2$.

Theorem 1.1. *We have*

$$\mathbb{E}[g(X)] = \sum_{k=1}^{\infty} g(x_k) \cdot p(x_k),$$

if the mean exists.

Example: Roll a fair die. Let $X = \#$ of dots on the side that comes up. Calculate $\mathbb{E}(X^2)$.

Solution: $\mathbb{E}(X^2) = \sum_{i=1}^{6} i^2 p(i) = 1^2 p(1) + 2^2 p(2) + 3^2 p(3) + 4^2 p(4) + 5^2 p(5) + 6^2 p(6) = \frac{1}{6} \cdot (1 + 4 + 9 + 16 + 25 + 36) = \frac{91}{6}$.

Calculate $\mathbb{E}(\sqrt{X}) = \sum_{i=1}^{6} \sqrt{i} p(i)$. Calculate $\mathbb{E}(e^X) = \sum_{i=1}^{6} e^i p(i)$. (Do at home)

Note that in general, $\mathbb{E}[g(X)] \neq g(\mathbb{E}(X))$.

$\mathbb{E}(X)$ is the expected value or 1-st *moment* of X. Then, $\mathbb{E}(X^n)$ is called the *n-th moment* of X.

1.2 Variance

We often seek to summarize the essential properties of a random variable in as simple terms as possible.

The mean is one such property. It gives a measure of the centre of the distribution.

Let $X = 0$ with probability 1.

Let $Y = \begin{cases} -2, & \text{with prob. } \frac{1}{3} \\ -1, & \text{with prob. } \frac{1}{6} \\ 1, & \text{with prob. } \frac{1}{6} \\ 2, & \text{with prob. } \frac{1}{3}. \end{cases}$

Both X and Y have the same expected value, but are quite different in other respects. One such respect is in their spread. We would like a measure of *spread.*

Definition: If X is a random variable with mean μ, then the *variance* of X, denoted by Var(X), is defined by

$$\mathrm{Var}(X) := \mathbb{E}[(X-\mu)^2] = \mathbb{E}(X^2) - [\mathbb{E}(X)]^2.$$

A small variance indicates a small spread.

Example: Roll a fair die. Let $X = \#$ that comes up. What is Var(X)?

Solution: Recall that $\mathbb{E}(X^2) = 91/6$, $\mathbb{E}(X) = (1+2+3+4+5+6)/6 = 21/6 = 7/2$. Thus,

$$\mathrm{Var}(X) = \frac{91}{6} - \left(\frac{7}{2}\right)^2 = \frac{91}{6} - \frac{7}{2} = \frac{182-147}{12} = \frac{35}{12}.$$

Proposition 1.2. *If a and b are constants then* $\mathrm{Var}(aX+b) = a^2\mathrm{Var}(X)$.

The square root of Var(X) is called the *standard deviation* of X, i.e.,

$$SD(X) = \sqrt{\mathrm{Var}(X)},$$

and it measures the scale of X.

1.3 Rules for Expectation and Variances

Suppose $X_1, X_2, \dots, X_n$ are random variables with means $\mu_1, \mu_2, \dots, \mu_n$, respectively, and variances $\sigma_1^2, \sigma_2^2, \dots, \sigma_n^2$, respectively.

What is the $\mathbb{E}(X_1 + X_2) = \mu_1 + \mu_2$. What about $\mathbb{E}(a_1X_1 + a_2X_2 + \ldots + a_nX_n)$?

$$\begin{aligned}\mathbb{E}(a_1X_1 + a_2X_2 + \ldots + a_nX_n) &= \mathbb{E}(a_1X_1) + \mathbb{E}(a_2X_2) + \ldots + \mathbb{E}(a_nX_n)\\ &= a_1\mathbb{E}(X_1) + a_2\mathbb{E}(X_2) + \ldots + a_n\mathbb{E}(X_n)\end{aligned}$$

Here we didn't care whether $X_1, \dots, X_n$ are *independent* or not. This is linearity of expectation.

Recall that $\mathrm{Var}\,(aX_1 + c) = a_1^2\mathrm{Var}\,(X_1)$. What can we say about similar variance calculations for sums of random variables?

If $X_1, \dots, X_n$ are **independent**, then we know that

$$\mathrm{Var}\,(X_1 + X_2) = \sigma_1^2 + \sigma_2^2 \qquad \text{and} \qquad \mathrm{Var}\,(X_1 - X_2) = \sigma_1^2 + \sigma_2^2.$$

What about $\mathrm{Var}\,(a_1X_1 + a_2X_2 + \ldots + a_nX_n)$?

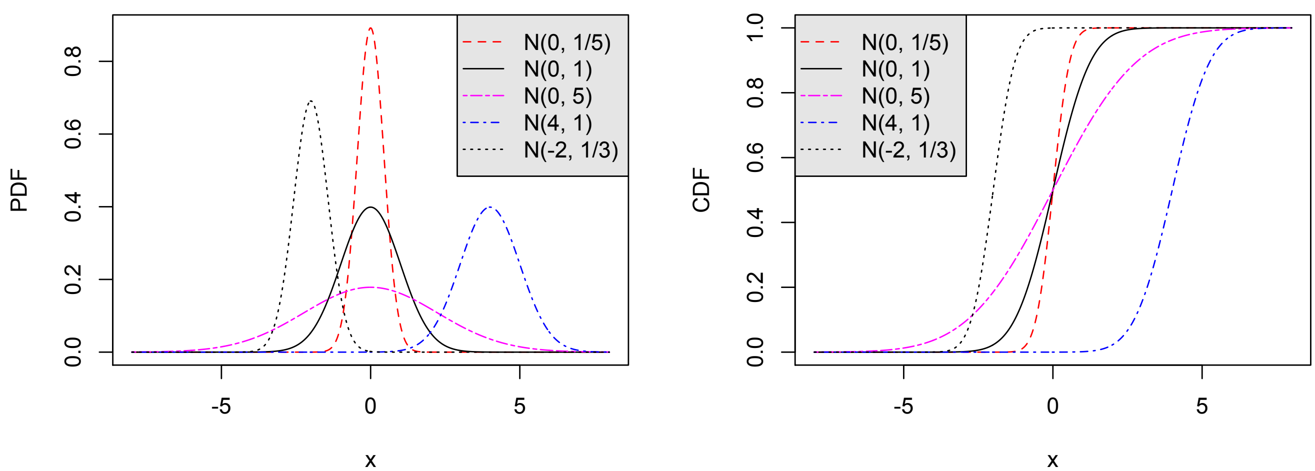
$$\begin{aligned}\mathrm{Var}\,(a_1X_1 + a_2X_2 + \ldots + a_nX_n) &= \mathrm{Var}\,(a_1X_1) + \mathrm{Var}\,(a_2X_2) + \ldots + \mathrm{Var}\,(a_nX_n)\\ &= a_1^2\mathrm{Var}\,(X_1) + a_2^2\mathrm{Var}\,(X_2) + \ldots + a_n^2\mathrm{Var}\,(X_n)\end{aligned}$$

1.4 Normal random variables

Definition: X is a called *normal* random variable with parameters μ and σ^2 (we write $X \sim N(\mu, \sigma^2)$) if the density of X is

$$\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(t-\mu)^2}{2\sigma^2}}.$$

Recall that $\mathbb{E}(X) = \mu$ and $\operatorname{Var}(X) = \sigma^2$.



Lemma 1.3. *If $X \sim N(\mu, \sigma^2)$, i.e., X is normally distributed with parameters μ and σ^2, then*

$$Y = aX + b \sim N(a\mu + b, a^2\sigma^2).$$

Definition: Z is a **standard** normal variable if it has a p.d.f.

$$f_Z(t) = \frac{1}{\sqrt{2\pi}} e^{-t^2/2}. \tag{2}$$

We say

$$Z \sim N(0, 1).$$

Proposition 1.4. *If X is normal random variable with parameters (μ, σ^2). Then the following empirical rules hold*

- *68% of the area lies between* $\mu - \sigma$ *and* $\mu + \sigma$
- *95% of the area lies between* $\mu - 2\sigma$ *and* $\mu + 2\sigma$
- *99.7% of the area lies between* $\mu - 3\sigma$ *and* $\mu + 3\sigma$*.*

Example: **How to read a Z-table?** If X is a normally distributed with parameters $\mu = 3$ and $\sigma^2 = 9$, find

(a) $\mathbb{P}(2 < X < 5)$

(b) $\mathbb{P}(X > 0)$

(c) $\mathbb{P}(|X - 3| > 6)$.

Solution:

(a) $\mathbb{P}(2 < X < 5) = \mathbb{P}(-\frac{1}{3} < Z < \frac{2}{3}) = \Phi(\frac{2}{3}) - \Phi(-\frac{1}{3}) \approx 0.3779$.

(b) $\mathbb{P}(X > 0) = \mathbb{P}(Z > -1) \approx 0.8413$.

(c) $\mathbb{P}(|X - 3| > 6) = \mathbb{P}(|Z| > 2) = \mathbb{P}(Z > 2) + \mathbb{P}(Z < -2) \approx 0.0456$.

Standard Normal Probabilities

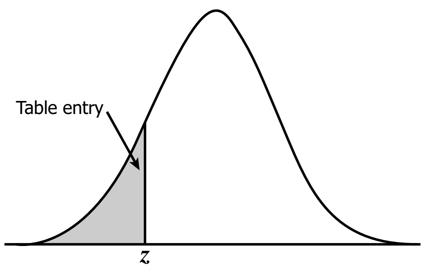


Table entry for z is the area under the standard normal curve to the left of z.

z	**.00**	**.01**	**.02**	**.03**	**.04**	**.05**	**.06**	**.07**	**.08**	**.09**
−3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
−3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
−3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
−3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
−3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
−2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
−2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
−2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
−2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
−2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
−2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
−2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
−2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
−2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
−2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
−1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
−1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
−1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
−1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
−1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
−1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681
−1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
−1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
−1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
−1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
−0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
−0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
−0.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
−0.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
−0.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
−0.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
−0.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
−0.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
−0.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
−0.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641

Standard Normal Probabilities

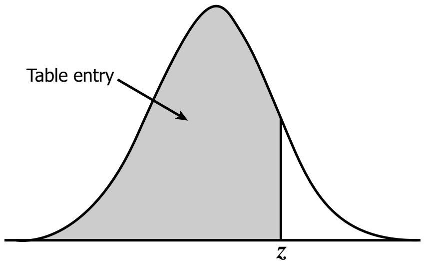


Table entry for z is the area under the standard normal curve to the left of z.

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998

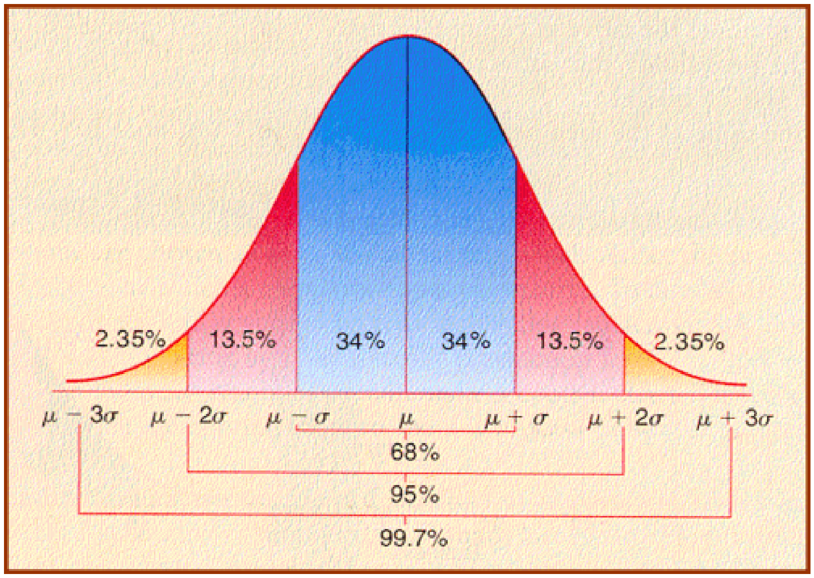


Figure 1: Empirical Rule for normal random variable with mean μ and variance σ^2.

Proposition 1.5. *If X and Y are independent normal random variables with parameters (μ_1, σ_1^2) and (μ_2, σ_2^2) respectively, then $X - bY$ is normal with mean $\mu_1 - b\mu_2$ and variance $\sigma_1^2 + b^2\sigma_2^2$.*

Definition: Suppose X is an observation from a population with mean μ and SD σ. The Z**-score** of X, usually denoted Z, is the number of SD's above (+) or below (−) the mean X is. Formula is

$$Z = \frac{X - \mu}{\sigma}.$$

The standard score indicates the relative standing of X in the population.

Calculating the standard score is a way of getting a common scale for different measurements which are approximately normally distributed.

Example: Suppose that a course has two midterms, for which the scores are approximately normally distributed, with means and SD's given below:

	Midterm 1	Midterm 2
Class Avg	55	50
Class SD	14	10
You get	76	67

On which test did you do better relative to the rest of the class?

2 Central Limit theorem

2.1 Properties of the sample mean

Suppose that $X_1, X_2, \ldots, X_n$ are n i.i.d r.v with mean μ and variance $\sigma^2 < \infty$. Let

$$\overline{X}_n := \frac{1}{n}(X_1 + \ldots + X_n) = \frac{1}{n}\sum_{i=1}^{n} X_i$$

be the sample average (or mean).

Theorem 2.1. $\mathbb{E}(\overline{X}_n) = \mu$ *and* $\mathrm{Var}(\overline{X}_n) = \sigma^2/n$.

Proof. Observe that

$$\mathbb{E}(\overline{X}_n) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(X_i) = \frac{1}{n}\cdot n\mu = \mu.$$

Also,

$$\mathrm{Var}(\overline{X}_n) = \frac{1}{n^2}\mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\cdot n\sigma^2 = \frac{\sigma^2}{n}.$$

□

See Figure 2 for histograms of $\overline{X}$ for different n.

2.2 Central Limit theorem

The WLLN says that $\overline{X}$ converges to μ. But it doesn't say how fast? Central Limit Theorem (CLT) finds the "right rate". WLLN talks about the convergence of the random variable while the CLT talks about the distribution of $\overline{X} - \mu$ and not the random variable.

A fundamental result in probability theory, is the CLT.

Theorem 2.2. *If $X_1, X_2, \ldots$ are i.i.d with mean μ and variance σ^2, then*

$$\overline{X} \overset{\mathcal{D}}{\sim} N(\mu, \sigma^2/n).$$

You can see the applet `http://onlinestatbook.com/stat_sim/sampling_dist/index.html` to see CLT in action.

Remark 2.1. *Note that the distribution of $\overline{X}$ is not affected by distribution of X except the fact that mean of $\overline{X}$ is the same as mean of X and Var$\left(\overline{X}\right) = \sigma^2/n$.*

Example 2.3. *Let $X_1, \ldots, X_n$ be i.i.d. $N(\mu, \sigma^2)$. What is the distribution of $\overline{X}$? What is the distribution of*

$$\frac{\overline{X} - \mu}{\sigma/\sqrt{n}}.$$

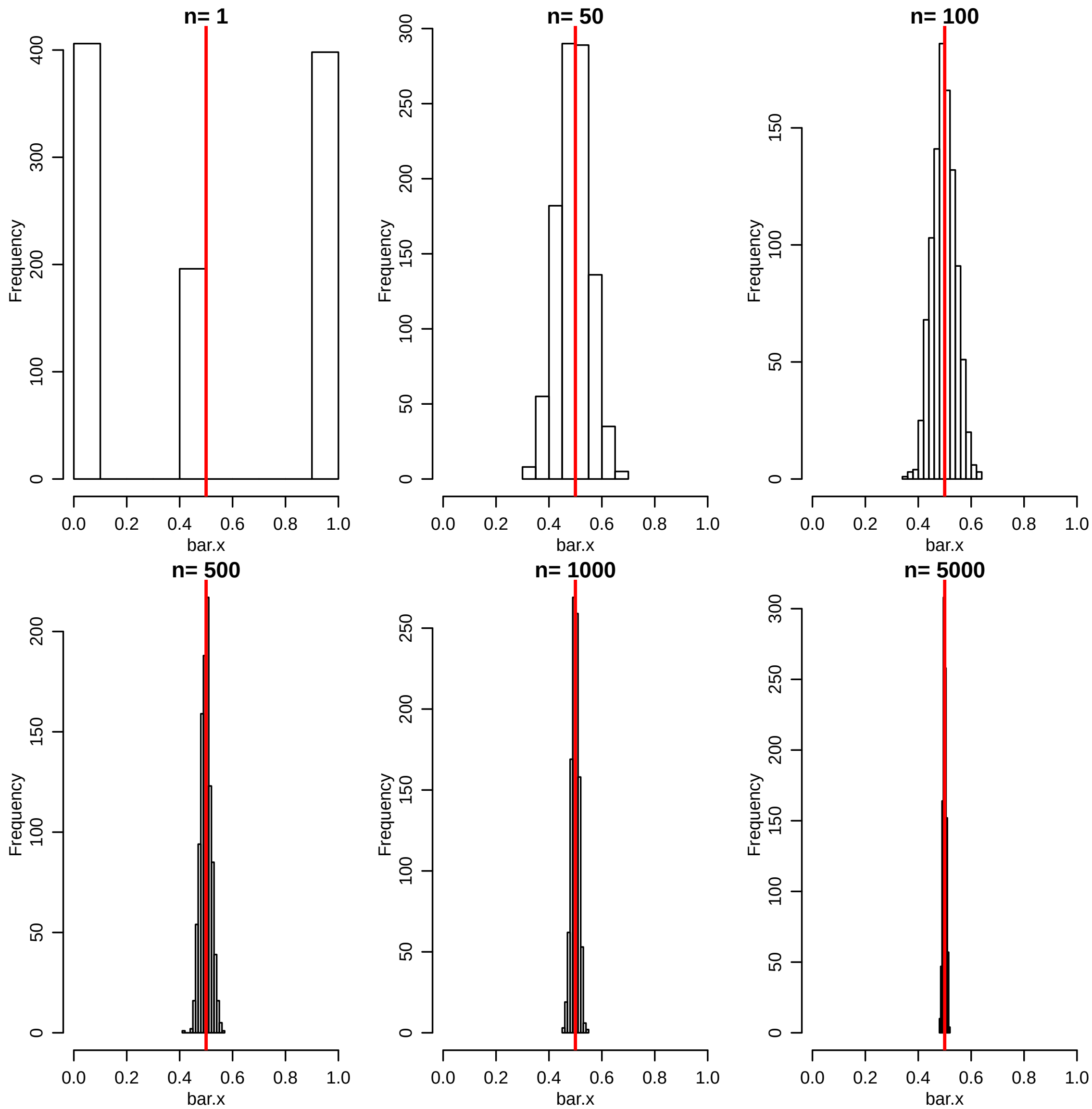


Figure 2: Histogram of $\overline{X}$ as n increases from 1 to 5000 when X_i takes values $0, .5$, and 1 with probabilities $.4, .2$, and $.4$, e.g., $\mathbb{P}(X = 1) = .4$. The red vertical lines represent the value of mean of X, which is equal to 0.5.

3 Confidence Intervals

Section A.6 (Page 1306) of the book

Let $A \leq B$ be two statistics that have the property that for all values of θ,

$$\mathbb{P}_\theta(A \leq \theta \leq B) = 1 - \alpha,$$

where $\alpha \in (0, 1)$. Then the **random** interval (A, B) is called an $(1 - \alpha) \times 100\%$ *confidence interval* for θ. See http://www.rossmanchance.com/applets/ConfSim.html for illustration

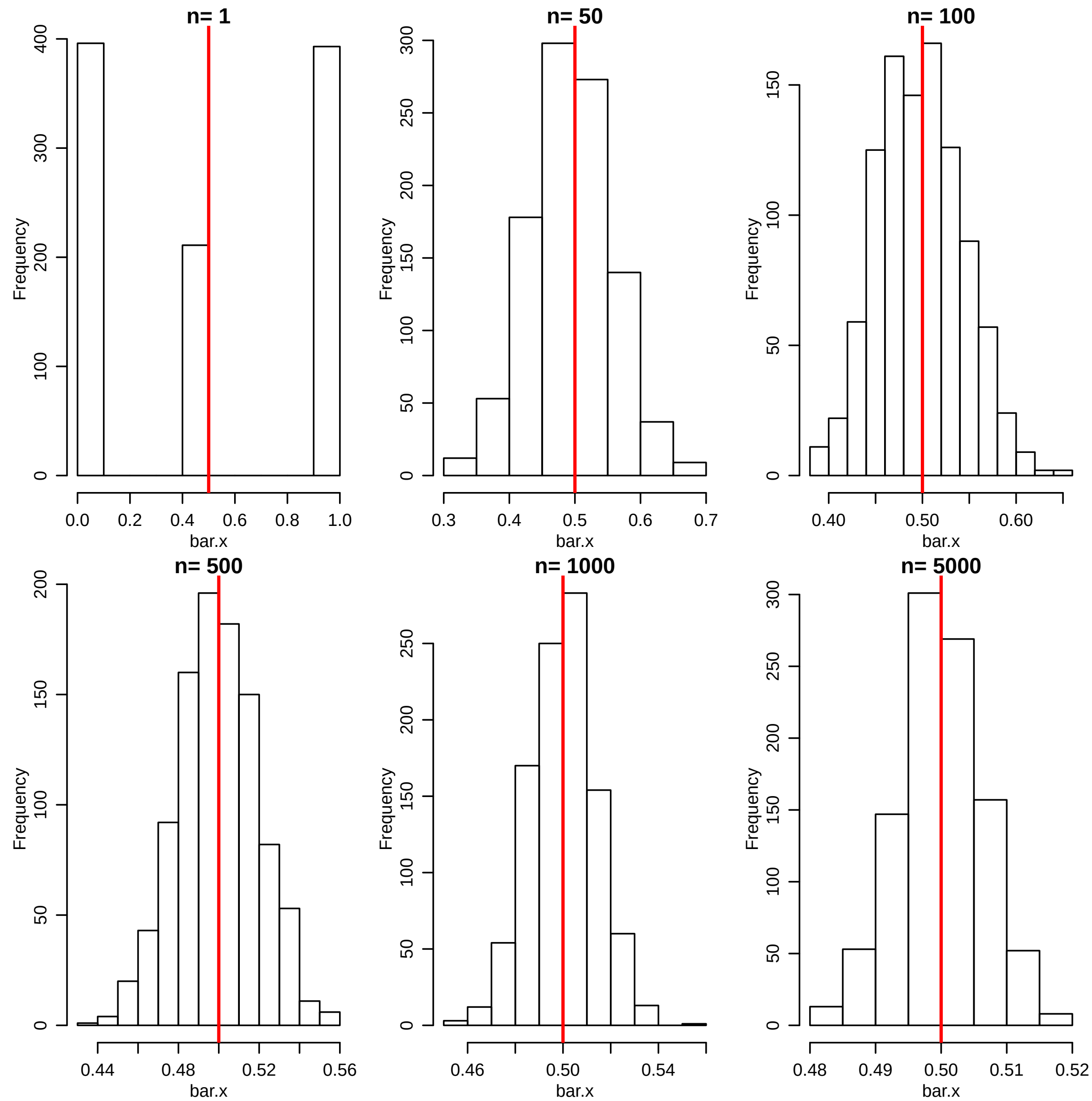


Figure 3: Histogram of $\overline{X}$. It's distribution is close to a Normal Random variable with mean μ and variance σ^2/n. Note that the x-axis is shrinking.

***Z*-confidence interval:** Suppose that $X_1, \ldots, X_n$ is a random sample with $\text{Var}(X) = \sigma^2$. Assume we know σ^2. We want to estimate $\mu = E(X)$.

Question: Construct a $(1-\alpha) \times 100\%$ confidence interval for μ.

Ans: Recall that

$$\overline{X} \overset{\mathcal{D}}{\sim} N(\mu, \sigma^2/n).$$

Thus

$$\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} o \overset{\mathcal{D}}{\sim} N(0, 1).$$

And our level $(1-\alpha)\times 100\%$ CI for μ is given by

$$\left[\overline{X}-\frac{\sigma}{\sqrt{n}}z_{1-\alpha/2},\overline{X}+\frac{\sigma}{\sqrt{n}}z_{1-\alpha/2}.\right]$$

(What is $z_{1-\alpha/2}$?) The above is a confidence interval because

$$\mathbb{P}\left(\overline{X}-\frac{\sigma}{\sqrt{n}}z_{1-\alpha/2}\leq\mu\leq\overline{X}+\frac{\sigma}{\sqrt{n}}z_{1-\alpha/2}\right)\approx 1-\alpha.$$

Exercise Suppose we have a $X_1,\dots,X_{100}$ such that $\mathrm{Var}(X)=25$. Suppose $\overline{X}=50$. Find a 95% confidence interval for μ.

t-confidence interval: What to do when $\mathrm{Var}(X)$ is unknown? If σ is unknown we will use S (the sample SD.)

What is S?

$$S^2 := \frac{1}{n-1}\sum_{i=1}^{n}(X_i-\overline{X})^2=\frac{1}{n-1}\left[(X_1-\overline{X})^2+\ldots+(X_n-\overline{X})^2\right]$$

And the confidence interval for μ is

$$\left[\overline{X}-\frac{S}{\sqrt{n}}t(1-\ \alpha/2;n-1),\overline{X}+\frac{S}{\sqrt{n}}t(1-\ \alpha/2;n-1)\right]$$

What is $t(1-\ \alpha/2,n-1)$? It is the $(1-\alpha/2)\times 100\%$ quantile of t-distribution with $n-1$ degrees of freedom.

Exercise Suppose $n=10$, $\overline{X}=20$, $S=4$. Find a 95% confidence interval for μ.

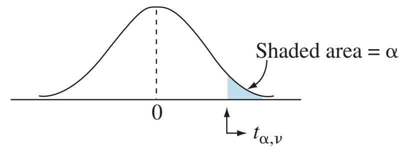


TABLE 2

Percentage points of Student's t distribution

df/α =	.40	.25	.10	.05	.025	.01	.005	.001	.0005
1	0.325	1.000	3.078	6.314	12.706	31.821	63.657	318.309	636.619
2	0.289	0.816	1.886	2.920	4.303	6.965	9.925	22.327	31.599
3	0.277	0.765	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	0.271	0.741	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	0.267	0.727	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	0.265	0.718	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	0.263	0.711	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	0.262	0.706	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	0.261	0.703	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	0.260	0.700	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	0.260	0.697	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	0.259	0.695	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	0.259	0.694	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	0.258	0.692	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	0.258	0.691	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	0.258	0.690	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	0.257	0.689	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	0.257	0.688	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	0.257	0.688	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	0.257	0.687	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	0.257	0.686	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	0.256	0.686	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	0.256	0.685	1.319	1.714	2.069	2.500	2.807	3.485	3.768
24	0.256	0.685	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	0.256	0.684	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	0.256	0.684	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	0.256	0.684	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	0.256	0.683	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	0.256	0.683	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	0.256	0.683	1.310	1.697	2.042	2.457	2.750	3.385	3.646
35	0.255	0.682	1.306	1.690	2.030	2.438	2.724	3.340	3.591
40	0.255	0.681	1.303	1.684	2.021	2.423	2.704	3.307	3.551
50	0.255	0.679	1.299	1.676	2.009	2.403	2.678	3.261	3.496
60	0.254	0.679	1.296	1.671	2.000	2.390	2.660	3.232	3.460
120	0.254	0.677	1.289	1.658	1.980	2.358	2.617	3.160	3.373
inf.	0.253	0.674	1.282	1.645	1.960	2.326	2.576	3.090	3.291

Source: Computed by M. Longnecker using Splus.

Exercise Suppose we have a $X_1 = 1$, $X_2 = -1$, $X_3 = 2$, $X_4 = 2$, $X_5 = 3$, $X_6 = -3$, $X_7 = 4$, $X_8 = -4$, $X_9 = 2$, and $X_{10} = -2$. Find a 95% confidence interval for μ. What does the probability mean?

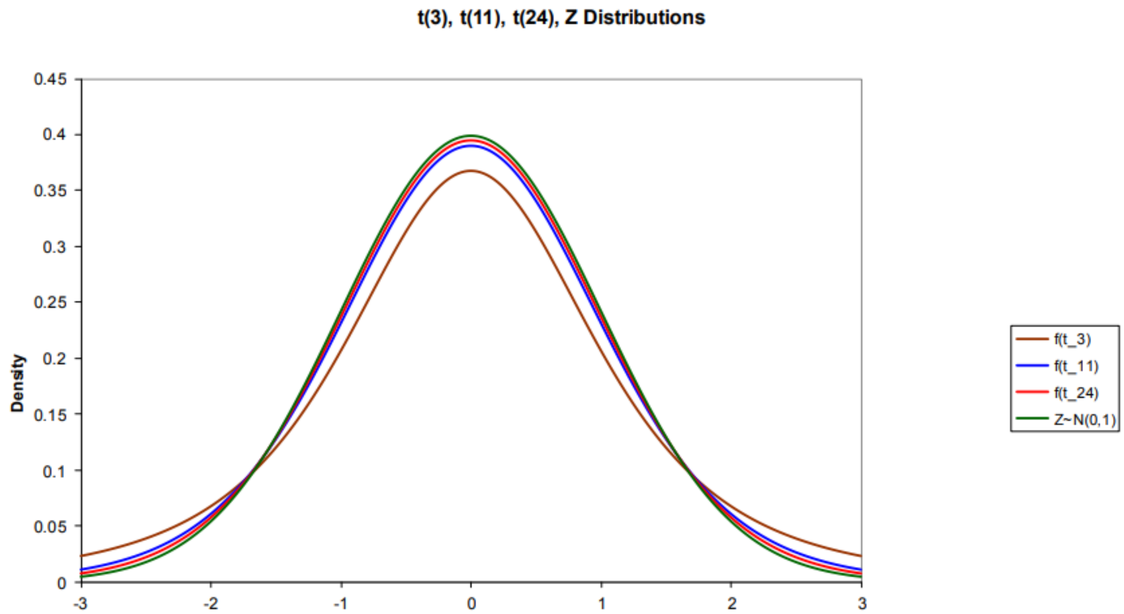


Figure 4: Students t density functions for different degrees of freedom. Comparison with standard normal density.

3.1 Comparison of means of two normal Samples (any sample sizes)

Suppose we have

1. The Y's are a random sample of size n_Y from a **normal** population with mean $\mu_Y$and SD σ_Y, both unknown; $\overline{Y}$ is the sample mean and $S_Y$is the sample SD.

2. The Z's are a random sample of size n_Z from a **normal** population with mean $\mu_Z$and SD σ_Z, both unknown; $\overline{Z}$ is the sample mean and $S_Z$is the sample SD.

3. The two samples are independent of one another.

4. The two standard deviations are equal; that is, $\sigma_Y = \sigma_Z$

A **confidence interval for $\mu_Y - \mu_Z$ is**

$$\left[\overline{Y} - \overline{Z} - S_{\overline{Y}-\overline{Z}}\, t(1-\ \alpha/2; n_Y + n_Z - 2), \right.$$
$$\left. \overline{Y} - \overline{Z} + \ S_{\overline{Y}-\overline{Z}}\, t(1-\ \alpha/2; n_Y + n_Z - 2)\right], \tag{3}$$

where

$$S^2_{\overline{Y}-\overline{Z}} = \sqrt{\frac{1}{n_Y} + \frac{1}{n_Z}}\ \frac{(n_Y - 1)S_Y^2 + (n_Z - 1)S_Z^2}{n_Y + n_Z - 2}$$

Example: A physician named a way of scoring a photograph of a face for wrinkles. The scores range from 1 to 20, with 1indicating no wrinkles, 20 indicating a severe case. She was trying to compare men who didn't use the anti-wrinkle cream to the group of men who used the anti-wrinkle cream. Let Y denote the scores for people who did **not** use the cream

$$n_Y = 10, \overline{Y} = 14, \sum_{i=1}^{10} (Y_i - \overline{Y})^2 = 105$$

Let Z denote the scores for people who did use the cream

$$n_Z = 20, \overline{Z} = 8, \sum_{i=1}^{20} (Z_i - \overline{Z})^2 = 224$$

Let us find a confidence interval for $\mu_Y - \mu_Z$.

Exercise: Compute

$$S^2_{\overline{Y}-\overline{Z}} = \sqrt{\frac{1}{10} + \frac{1}{20}}\frac{105 + 224}{10 + 20 - 2} = 1.7625 \quad \text{and} \quad t(.975, 28) = 2.048$$

Hence the 95% confidence interval is

$$[(14 - 8) - 2.048 \times 1.328, (14 - 8) + 2.048 \times 1.328] = [3.3, 8.7]$$

4 Correlation

4.1 Correlation coefficient

Definition: The *correlation* between X and Y, denoted by $\rho(X, Y)$, is defined (as long as $\mathrm{Var}(X)$ and $\mathrm{Var}(Y)$ are positive) by

$$\rho(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}}.$$

It can be shown that

$$-1 \leq \rho(X,Y) \leq 1,$$

with equality if and only if $Y = aX + b$ (assuming $\mathbb{E}(X^2)$ and $\mathbb{E}(Y^2)$ are both finite).

The correlation coefficient is therefore a measure of the degree of *linear association* between X and Y. If $\rho(X,Y) = 0$ then this indicates no linearity, and X and Y are said to be *uncorrelated*.

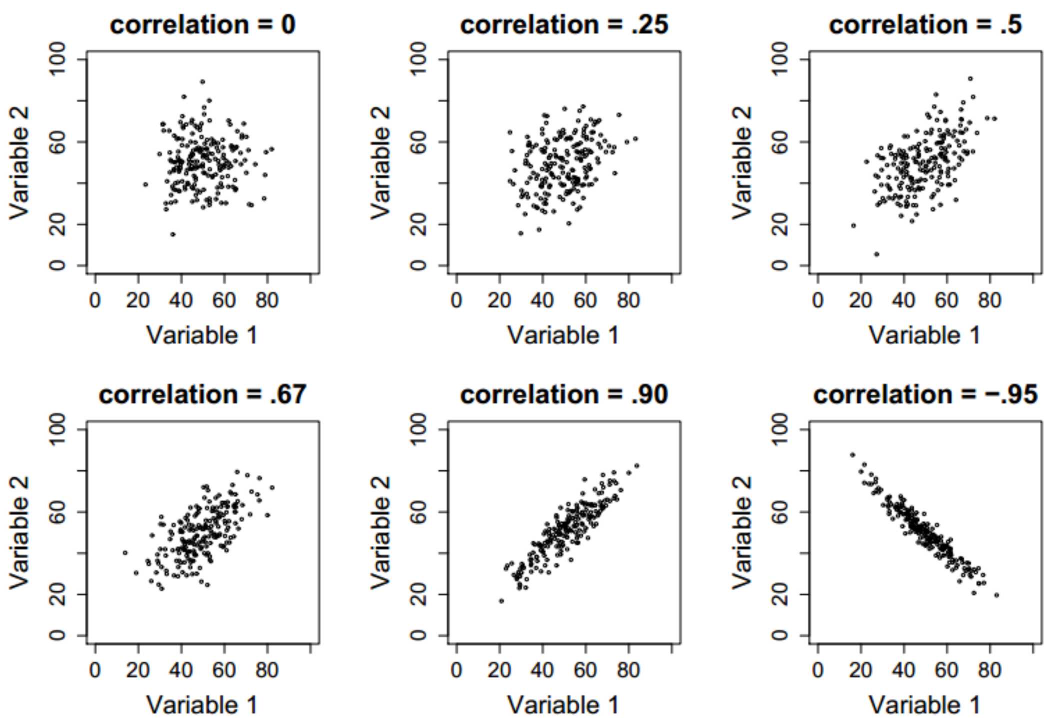


Figure 5: Figure showing examples of correlation

But it is important to stress that if correlation is zero it **does not** imply that there is **no** relationship. It only implies no linear relationship exists.

If you want to see whether there is a relationship between two variables, should always look at the scatter plot.

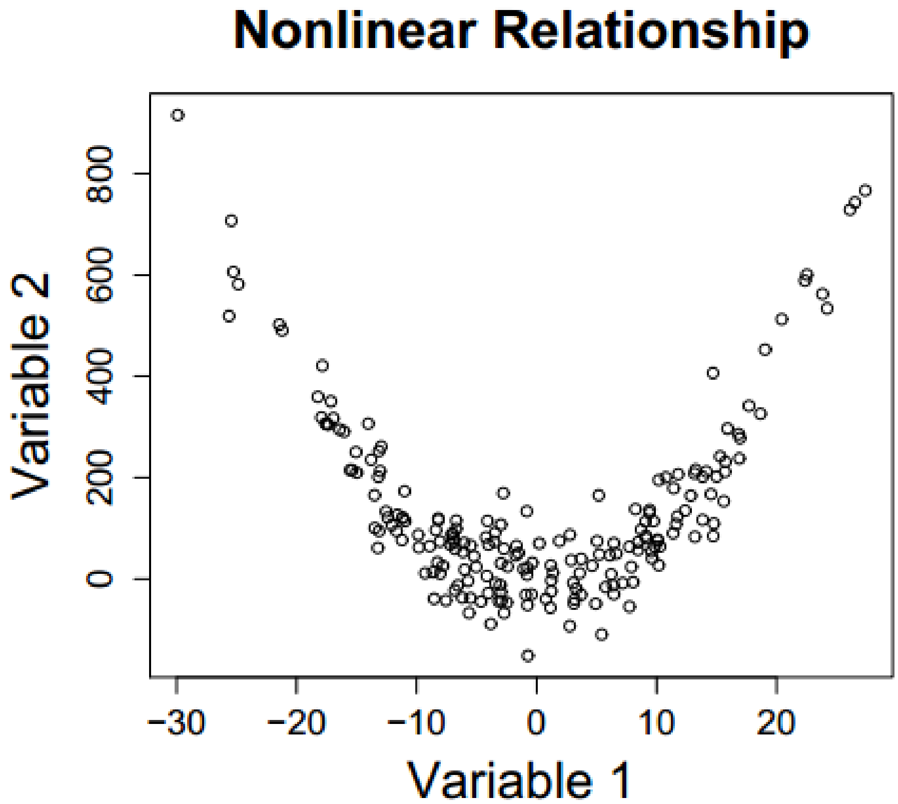


Figure 6: Beware of interpretation of correlation

On the surface correlations seems to be a very useful quantity. However, there are many shortcomings of the correlation.