Confidence Intervals

Points → Intervals

Distribution or population

Estimate parameters

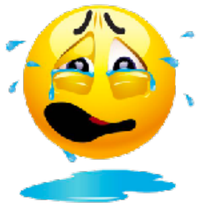
← Point estimates

$\mu \approx 3.14$

$p \approx 0.48$

Precise

Certainly wrong



No confidence

→ Confidence intervals

Precision ↘

Confidence ↗

With 95% confidence (probability) $\mu \in (3.1, 3.18)$

Back to Normal

CLT

Averages normally distributed

Intuition

Almost everything

$$X_1, \ldots, X_n \perp\!\!\!\perp, \sim \text{ any distribution with mean } \mu, \text{ and stdv } \sigma$$

$$\overline{X^n} \stackrel{\text{def}}{=} \frac{X_1 + \ldots + X_n}{n}$$

Sample mean

$$Z_n \stackrel{\text{def}}{=} \frac{(X_1 + \ldots + X_n) - n\mu}{\sigma\sqrt{n}}$$

Typically ≥ 30

Roughly

CLT

For sufficiently large n

$$Z_n \dot{\sim} \mathcal{N}(0, 1)$$

Standard Normal Variable

Predicting Standard Normal

Standard Normal Variable

Predict value of Z

Point prediction

Highest probability

Unbiased

Precise

Wrong

Interval

$-a \leq Z \leq a$

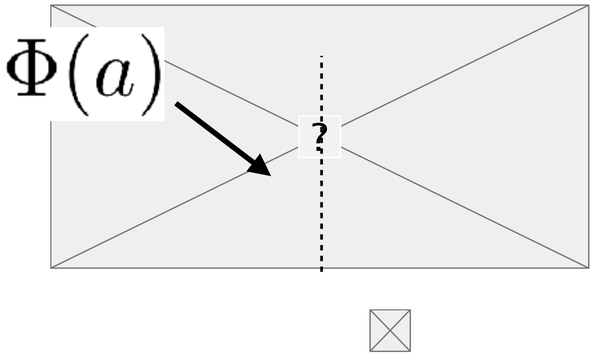
$P(-a \leq Z \leq a) > 0$

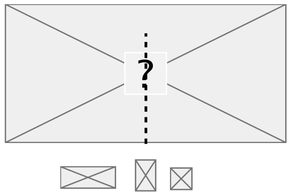
P = ?

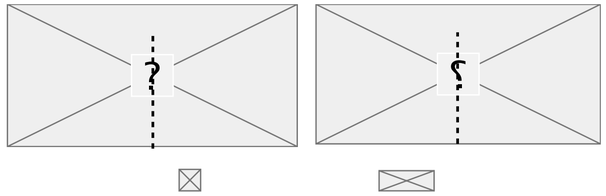
Interval Probability

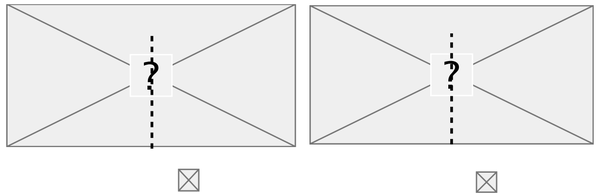
$$Z \sim \mathcal{N}(0,1)$$

$$\Phi(a) = F(a) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{a} e^{-t^2/2} dt$$





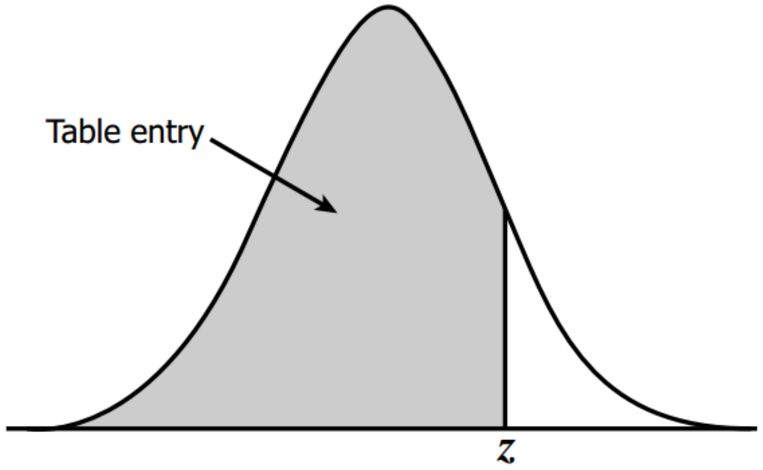




Calculating $\Phi(a)$

No known formula

Use Z / Standard Normal Table



$$\Phi(1) = 0.8413$$

Program

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

Python

In scipy.stats　norm.cdf(x)　cumulative distribution function

$\Phi(1)$

```
from scipy.stats import norm
norm.cdf(1)
0.841344746069
```

$\Phi(2)$

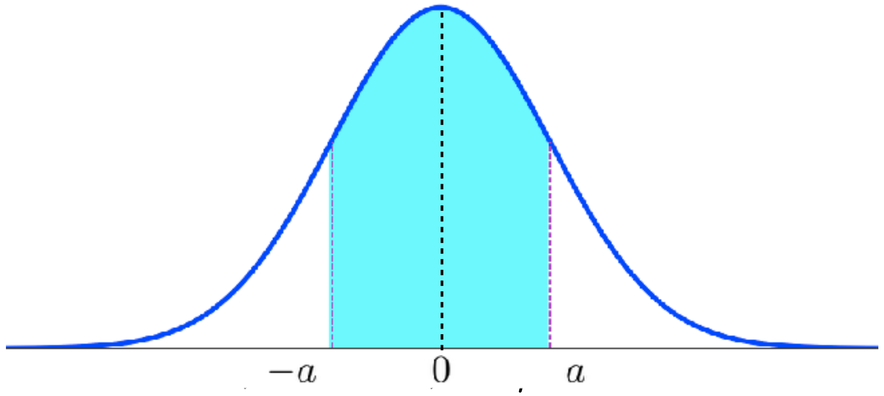
```
norm.cdf(2)
0.977249868052
```

$\Phi(3)$

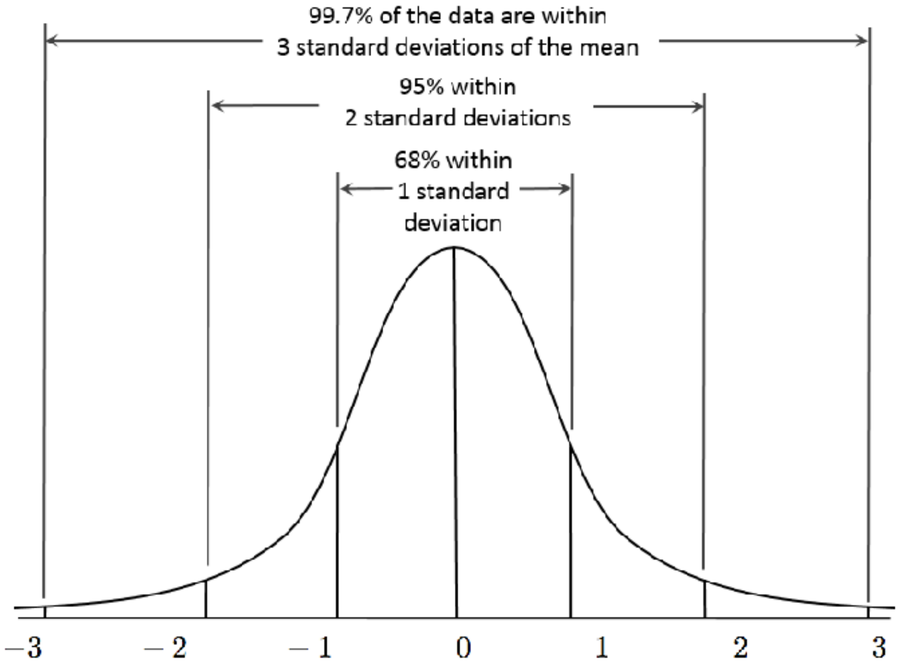
```
norm.cdf(3)
0.998650101968
```

68 - 95 - 99.7 Rule

$$P(-a \leq Z \leq a) = 2\Phi(a) - 1$$



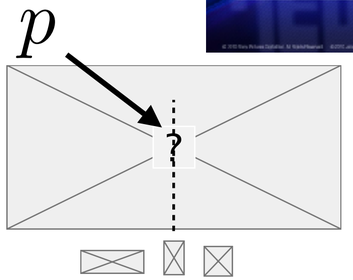
a	P(-a ≤ Z ≤ a)
1	2 · 0.8413 - 1 = 0.682
2	2 · 0.9772 - 1 = 0.9544
3	2 · 0.9987 - 1 = 0.9974



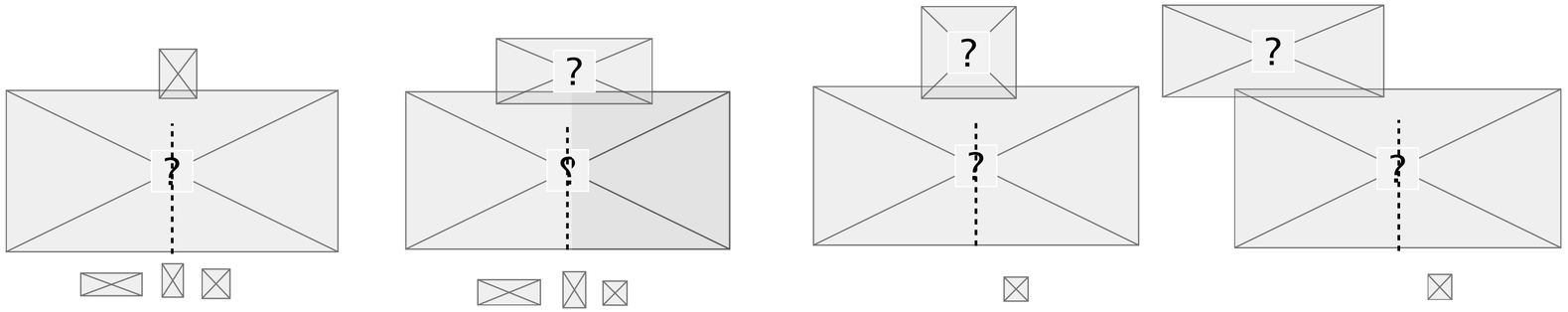
Interval → Probability

Typically Given desired probability p

Find a s.t. $P(-a \leq Z \leq a) = p$



Saw →



Python

norm.ppf(p)

percent point function

converts percentile to a point

$\Phi^{-1}(0.95)$

```
from scipy.stats import norm
norm.ppf(0.95)
1.64485362695
```

$\Phi^{-1}(0.975)$

```
norm.ppf(0.975)
1.95996398454
```

$\Phi^{-1}(0.99)$

```
norm.ppf(0.99)
2.32634787404
```

Common Values

p=95%

```
norm.ppf(0.975)
1.95996398454
```

$$a = \Phi^{-1}\left(\frac{1+p}{2}\right) = \Phi^{-1}(0.975) \approx 1.96$$

$$P(-1.96 \leq Z \leq 1.96) \approx 0.95$$

68 - 95 - 99.7

$$P(-2 \leq Z \leq 2) \approx 0.95$$

p	$\frac{1+p}{2}$	$\Phi^{-1}\left(\frac{1+p}{2}\right)$
90	0.95	1.645
95	0.975	1.960
98	0.99	2.056

General Normal Distributions

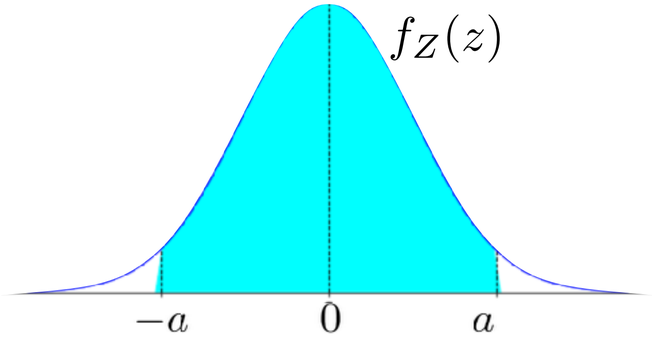
$X \sim \mathcal{N}_{\mu,\sigma^2} \qquad \mathcal{N}(\mu, \sigma^2)$

X, Z

normal

$Z \overset{\text{def}}{=} \frac{X-\mu}{\sigma} \sim \mathcal{N}_{0,1}$

Standard Normal



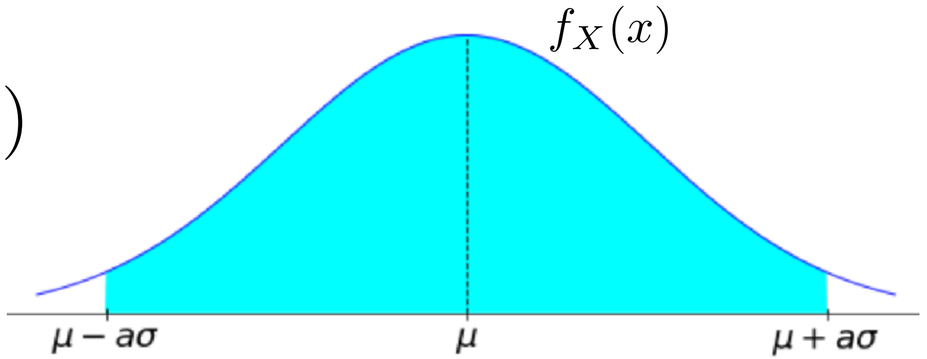
$$P(\mu - a\sigma \leq X \leq \mu + a\sigma) = P(-a\sigma \leq X - \mu \leq a\sigma)$$

X within “a” std from its mean

$$= P(-a \leq \frac{X-\mu}{\sigma} \leq a)$$

$$= P(-a \leq Z \leq a)$$

Z within “a” std from its mean



Example

X ~N(1, 4)

p = 0.95

μ = 1

σ = 2

Z within 1.96 std from its mean

X within 1.96 std from its mean

$$0.95 \approx P(-1.96 \leq Z \leq 1.96) = P(\mu - 1.96\sigma \leq X \leq \mu + 1.96\sigma)$$

$$= P(-2.92 \leq X \leq 4.92)$$

Confidence Intervals

Any parameter

Simplest and by far most common

mean μ

proportion p

Given a sample X_1, … Xn

Find an interval containing μ

First

σ known

Next lecture

σ unknown

Sample-Mean Distribution

$$\frac{X_1 + ... + X_n - n\mu}{\sigma\sqrt{n}} \dot{\sim} \mathcal{N}(0,1) \quad \times \frac{\sigma}{\sqrt{n}}$$

$$\frac{X_1 + ... + X_n - n\mu}{n} \dot{\sim} \mathcal{N}\left(0, \frac{\sigma^2}{n}\right) \quad +\mu$$

$$\overline{X} = \frac{X_1 + ... + X_n}{n} \dot{\sim} \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$$

Sampling Distribution of the sample mean

$\overline{X}$

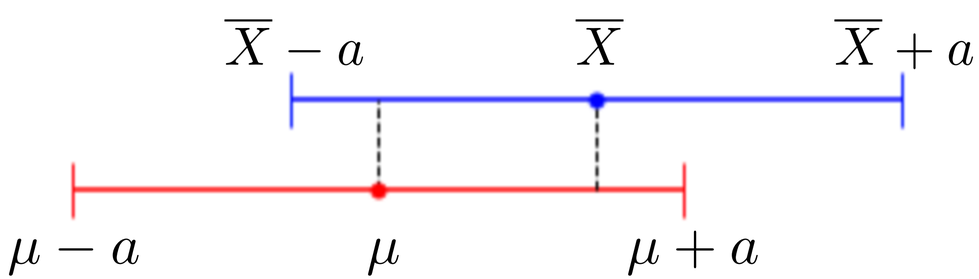
- Roughly normal
- Centered at sample mean $\mu_{\overline{X}} = \mu$
- Standard deviation $V(\overline{X}) = \frac{\sigma^2}{n} \quad \sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}$

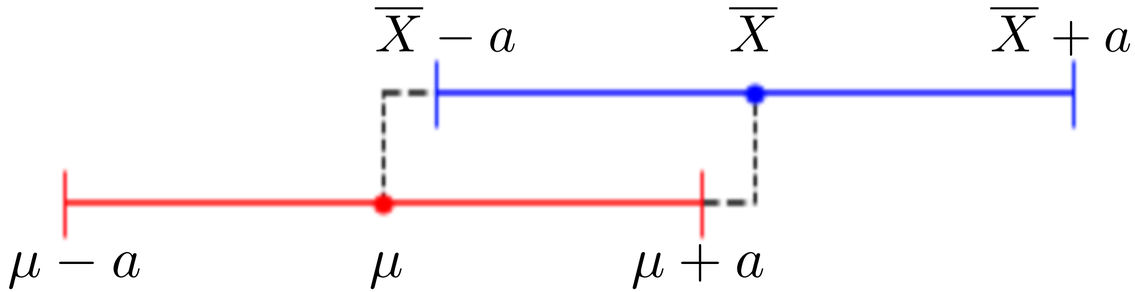
Standard error

Proximity is Reciprocal

← $\overline{X}$ near μ → μ near $\overline{X}$

$\overline{X} \in (\mu - a, \mu + a)$ $|\overline{X} - \mu| < a$ $\mu \in (\overline{X} - a, \overline{X} + a)$

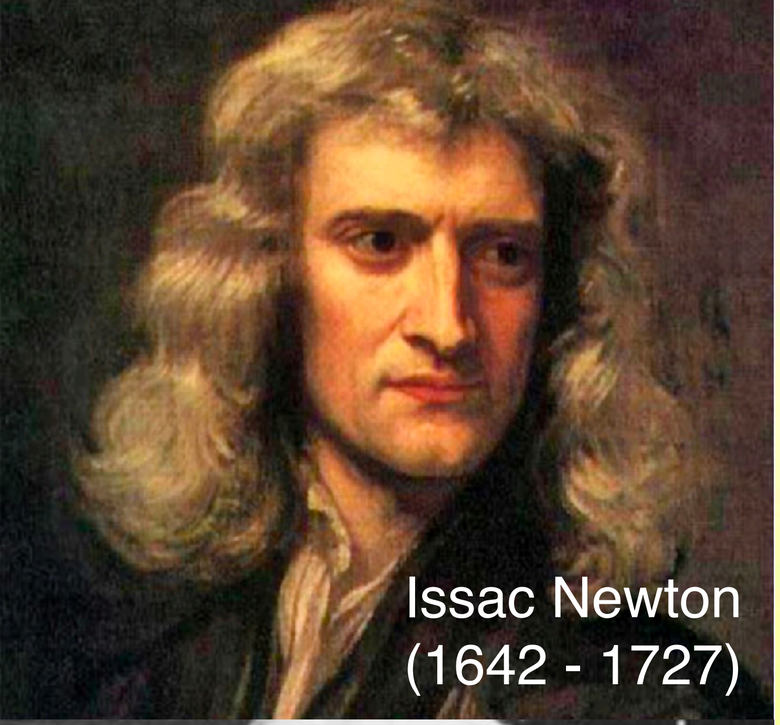




$\overline{X} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$

With high probability $\overline{X}$ near μ

With high probability μ near $\overline{X}$



Issac Newton
(1642 - 1727)

Jerzy Neyman
(1894 - 1981)

The apple doesn’t fall far from the tree

The tree doesn’t grow far from the apple

Confidence Interval

With probability p

$$\overline{X} \in (\mu - z_p \sigma_{\overline{X}}, \mu + z_p \sigma_{\overline{X}})$$

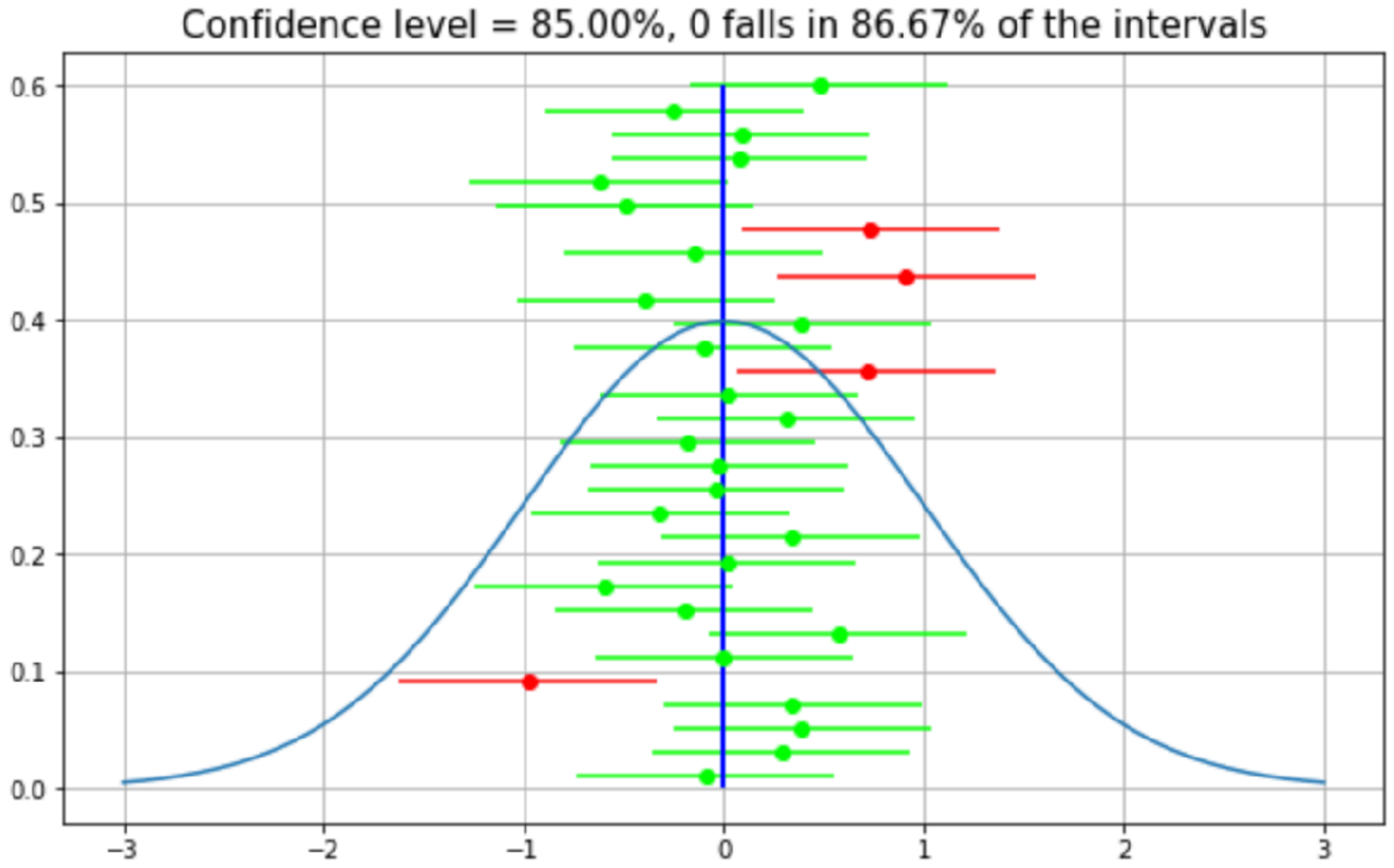
$$\left|\overline{X} - \mu\right| < z_p \sigma_{\overline{X}}$$

$$\mu \in (\overline{X} - z_p \frac{\sigma}{\sqrt{n}}, \overline{X} + z_p \frac{\sigma}{\sqrt{n}})$$

sample size: 5

number of experime--nts: 30

confiden--ce: 0.85



Daily Tweets

tweets of a random Tweeter user is a random variable with σ=2

In a sample of 121 users the sample mean was 3.7

Find the 95% confidence interval for the distributiontion mean

$$z_p = \Phi^{-1}(\frac{1+p}{2}) = \Phi^{-1}(0.975) = 1.96$$

95% confidence interval for mean

$$(\overline{X} - z_p\sigma_{\overline{X}}, \overline{X} + z_p\sigma_{\overline{X}}) = (\overline{X} - z_p\frac{\sigma}{\sqrt{n}}, \overline{X} + z_p\frac{\sigma}{\sqrt{n}})$$

$$= (3.344, 4.056)$$

Margin of error

Heart Rate per Minute

Adult heart rate has standard deviation σ=7.5 beats per minute

Estimate average heart rate within margin of error < 2

With confidence level 90%

$$z_p = \Phi^{-1}(\frac{1+p}{2}) = \Phi^{-1}(0.95) = 1.645$$

Sample size

$$z_p \sigma_{\overline{X}} = z_p \frac{\sigma}{\sqrt{n}} = 2 \qquad n = \left(z_p \frac{\sigma}{2} \right)^2 = 38.05$$

Confidence Intervals