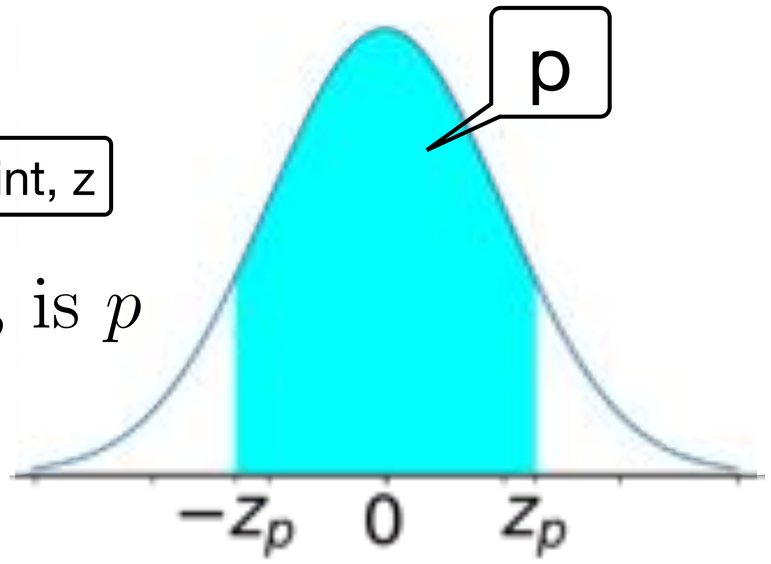
Confidence Intervals
Unknown σ

- Estimate σ
- Student's t-distribution
- Step-by-step instructions
- Example

Confidence Intervals - Known σ

Standard normal distribution $\mathcal{N}_{0,1}$



$0 \le p \le 1 \qquad z_p : \text{ point s.t. area between } -z_p \text{ and } z_p \text{ is } p$

$$Z \sim \mathcal{N}_{0,1} \quad p = P(|Z| \le z_p) = P(-z_p \le Z \le z_p)$$

$$= \Phi(z_p) - \Phi(-z_p) = \Phi(z_p) - (1 - \Phi(z_p)) = 2\Phi(z_p) - 1$$

CDF of standard normal

$$\Phi(z_p) = \tfrac{1+p}{2}$$

$$z_p = \Phi^{-1}\left(\tfrac{1+p}{2}\right)$$

$$z_{0.9} = \Phi^{-1}\left(\tfrac{1+0.9}{2}\right) = \Phi^{-1}(0.95) = 1.645$$

$$P(|Z| \le 1.645) = 0.9$$

Sample Mean ≈ Normal

$X_1, X_2, \ldots, X_n$ i.i.d. known σ unknown μ

$\overline{X} = \frac{X_1 + \ldots + X_n}{n}$ Sample mean

$\mu_{\overline{X}} = \mu$ Unbiased

$V(\overline{X}) = \frac{\sigma^2}{n}$ $\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}$

$$\frac{\overline{X} - \mu}{\sigma_{\overline{X}}} = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \dot{\sim} N_{0,1}$$

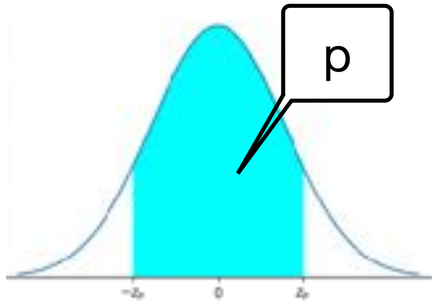
mean 0

std 1

CLT

Confidence Interval

Standard normal



$$Z \sim \mathcal{N}_{0,1} \qquad P(|Z| \leq z_p) = p$$

$$\frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \dot{\sim} \mathcal{N}_{0,1}$$

$$P\left(\left|\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}\right| \leq z_p\right) \approx p$$

$$P\left(\left|\overline{X}-\mu\right| \leq z_p \frac{\sigma}{\sqrt{n}}\right) \approx p$$

With probability ≈ p

$$\left|\overline{X}-\mu\right| \leq z_p \frac{\sigma}{\sqrt{n}}$$

Margin of error

$$\mu \in \left[\overline{X} - z_p \frac{\sigma}{\sqrt{n}}, \overline{X} + z_p \frac{\sigma}{\sqrt{n}}\right]$$

Confidence → Interval

Given confidence p samples $X_1, \ldots, X_n$

Determine Critical value (z) $z_p = \Phi^{-1}\left(\frac{1+p}{2}\right)$ `norm.ppf((1+p)/2)`

Sample mean $\overline{X} = \frac{X_1+\ldots+X_n}{n}$

Margin of error $z_p \frac{\sigma}{\sqrt{n}}$ σ known

Confidence interval $\left[\overline{X} - z_p \frac{\sigma}{\sqrt{n}}, \overline{X} + z_p \frac{\sigma}{\sqrt{n}}\right]$

Problem? σ almost never known

Unknown σ

$X_1, X_2, \ldots, X_n \quad \perp\!\!\!\perp \quad \mathcal{N}_{\mu,\sigma}$ Neither σ nor μ known μ = ?

$\overline{X} = \frac{X_1+\ldots+X_n}{n}$ Sample mean $\mu_{\overline{X}} = \mu$ Unbiased $\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}$

$\frac{\overline{X}-\mu}{\sigma_{\overline{X}}} = \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \dot{\sim} \mathcal{N}_{0,1}$ Standard normal

$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$ Sample variance Bessel corrected

$\mu_{S^2} = \sigma^2$ Unbiased

$\frac{\overline{X}-\mu}{S/\sqrt{n}}$ Almost standard $S \approx \sigma$ normal $S - r.v.$

Student's t-distribution n-1 degrees of freedom

Student’s t-distribution

$$T_\nu = \frac{\overline{X}-\mu}{S/\sqrt{\nu+1}}$$

Student’s t-distribution, ν degrees of freedom

PDF

$$T_\nu \sim f_\nu(t) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\cdot\Gamma(\frac{\nu}{2})}\left(1+\frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}}$$

Gamma function

Only dependence on t

Symmetric around 0

See a bit more

t object

in scipy.stats module

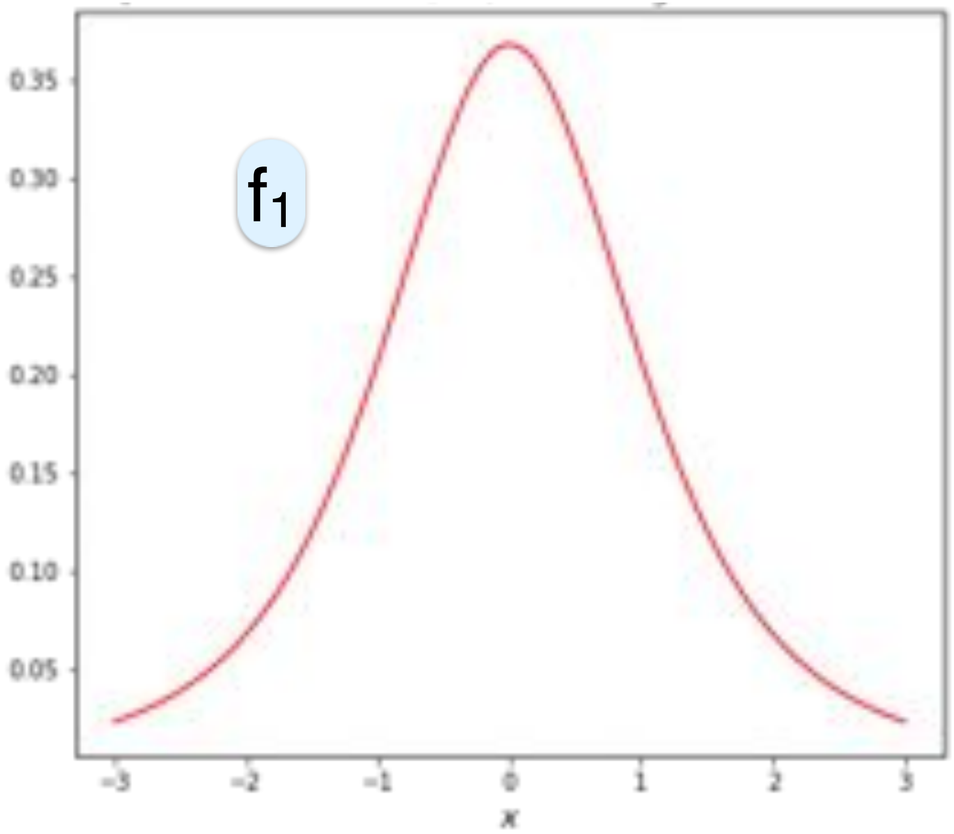
Degrees of freedom

probability density function

t.pdf(x,ν)

$f_3(1)$

```
from scipy.stats import t
t.pdf(1,3)
0.20674833578317203
```

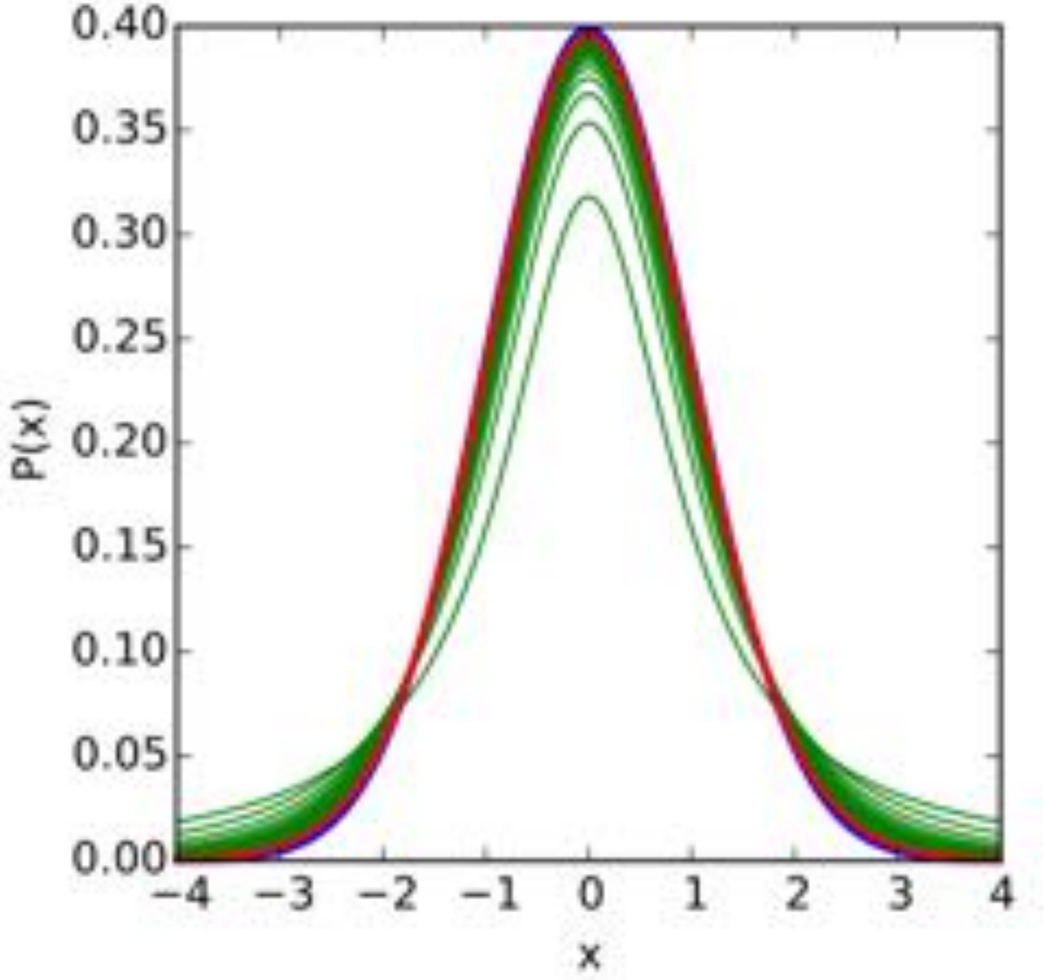


Bell shaped

Similar to Gaussian

Dependence on ν

As ν increases $f_\nu(t) \to \phi(t)$



standard normal distribution

t-distribution ν=30

Logical

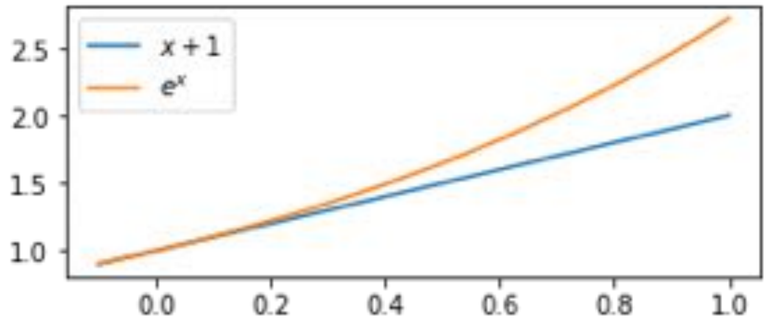
S → constant σ

Examples in notebook

Analytical Argument

$$f_\nu(t) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\cdot\Gamma(\frac{\nu}{2})}\left(1+\frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}}$$

For small x $\quad 1+x \sim e^x$



As $\nu\to\infty$ $\quad \left(1+\frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}} \approx \left(e^{\frac{t^2}{\nu}}\right)^{-\frac{\nu+1}{2}} \approx e^{-\frac{t^2}{\nu}\frac{\nu+1}{2}} \approx e^{-\frac{t^2}{2}}$ Standard Normal

Distribution $\quad$ Constant same $\quad \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\Gamma(\frac{\nu}{2})} \approx \frac{1}{\sqrt{2\pi}}$

William Sealy Gosset

Guinness Brewery

World's largest

Billion pints / year

~1.75 Billion bottles

Quality

Consistency

Statisticians

Trade secret

Publish

Pseudonym

VOLUME VI MARCH, 1908 No. 1

BIOMETRIKA.

THE PROBABLE ERROR OF A MEAN.

BY STUDENT.

Introduction.

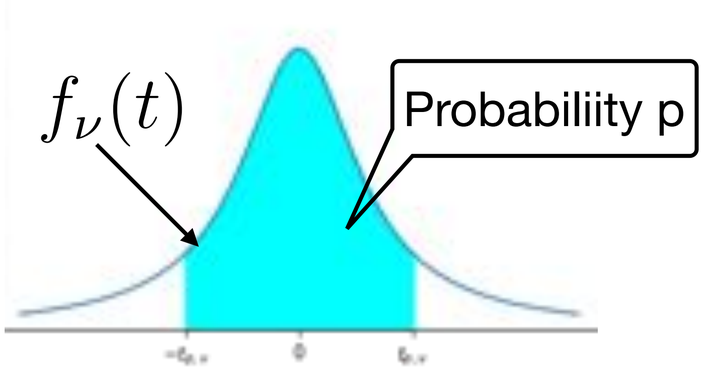
ANY experiment may be regarded as forming an individual of a "population" of experiments which might be performed under the same conditions. A series of experiments is a sample drawn from this population.

Déjà vu

T_ν Student's t-distribution, ν degrees of freedom

Critical value, t

$t_{p,\nu}: \quad P(|T_\nu| \leq t_{p,\nu}) = p$

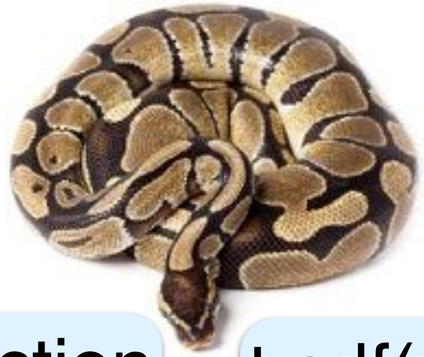


$$p = P(|T_\nu| \leq t_{p,\nu}) = P(-t_{p,\nu} \leq T_\nu \leq t_{p,\nu}) = 2F(t_{p,\nu}) - 1$$

CDF of T_ν

$$F(t_{p,\nu}) = \frac{1+p}{2}$$

$$t_{p,\nu} = F_\nu^{-1}\left(\frac{1+p}{2}\right)$$



Degrees of freedom

Cumulative distribution function

t.cdf(x,ν)

$F_3(1)$

```
from scipy.stats import t
t.cdf(1,3)
0.8044988905221147 6
```

Inverse cdf

percent point function

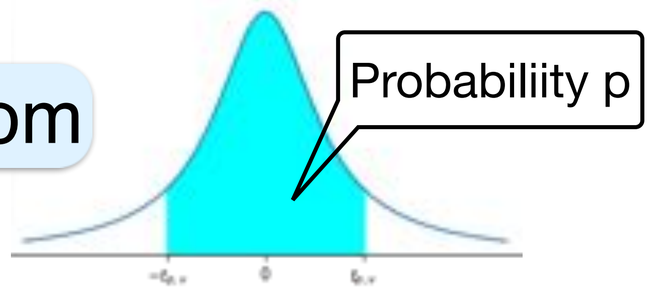
t.ppf(x,ν)

$F_3^{-1}(0.95)$

```
t.ppf(0.95,3)
2.353363434801826 4
```

Confidence Interval

t-distribution, ν degrees of freedom



t-statistic

$$\frac{\overline{X}-\mu}{S/\sqrt{n}} \sim f_{n-1}(t)$$

$$P\left(\left|\frac{\overline{X}-\mu}{S/\sqrt{n}}\right| \leq t_{p,n-1}\right) = p$$

$$P\left(\left|\overline{X}-\mu\right| \leq t_{p,n-1}\frac{S}{\sqrt{n}}\right) = p$$

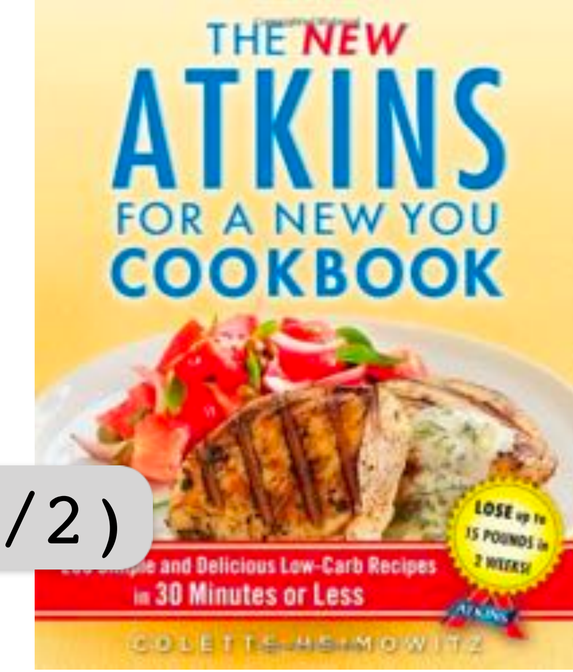
With probability p

margin of error

$$\left|\overline{X}-\mu\right| \leq t_{p,n-1}\frac{S}{\sqrt{n}}$$

$$\mu \in \left[\overline{X} - t_{p,n-1}\frac{S}{\sqrt{n}}, \overline{X} + t_{p,n-1}\frac{S}{\sqrt{n}}\right]$$

Confidence → Interval



Given confidence p samples $X_1, \ldots, X_n$

Determine critical t $t_{p,n-1} = F_{n-1}^{-1}\left(\frac{1+p}{2}\right)$ `t.ppf((1+p)/2)`

Sample mean $\overline{X} = \frac{X_1+\ldots+X_n}{n}$

Sample variance $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$

Margin of error $t_{p,n-1}\frac{S}{\sqrt{n}}$ σ unnecessary

Confidence interval $\left[\overline{X} - t_{p,n-1}\frac{S}{\sqrt{n}}, \overline{X} + t_{p,n-1}\frac{S}{\sqrt{n}}\right]$

Mature African Elephant Trunk Length

Mature African Elephant Trunk Length

8 measurements 5.62, 6.07, 6.64, 5.91, 6.30, 6.55, 6.19, 5.48 feet

Find 95% confidence interval for distribution mean

Critical t $t_{p,n-1} = F_{n-1}^{-1}\left(\frac{1+p}{2}\right) = F_7^{-1}(0.975) \approx 2.3646$

```
t.ppf(0.975,7)
2.3646
```

Sample mean $\overline{X} = 6.095$

Sample variance $S^2 \approx 0.1705$ $S = 0.4130$

Margin of error $t_{p,n-1}\frac{S}{\sqrt{n}} \approx 0.3453$

Confidence interval $\left(\overline{X} - t_{p,n-1}\frac{S}{\sqrt{n}}, \overline{X} + t_{p,n-1}\frac{S}{\sqrt{n}}\right)$

$\approx (5.7497, 6.4403)$

Observations

n large

$f_{n-1}(t) \to \phi(t)$

$t_{p,n-1} \to z_p$

S →

σ

} Can use z-based techniques

n small

t-distribution more accurate

Yields larger margin of error than known σ

Assumed $X_i \sim N$, best when this roughly holds

Confidence Intervals
Unknown σ

- Estimate σ
- Student's t-distribution
- Step-by-step instructions
- Example

Confidence Intervals

Points → Intervals

Distribution or population

Estimate parameters

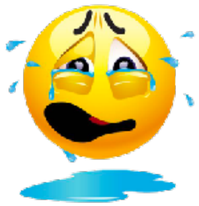
← Point estimates

$\mu \approx 3.14$

$p \approx 0.48$

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

$-a \leq Z \leq a$

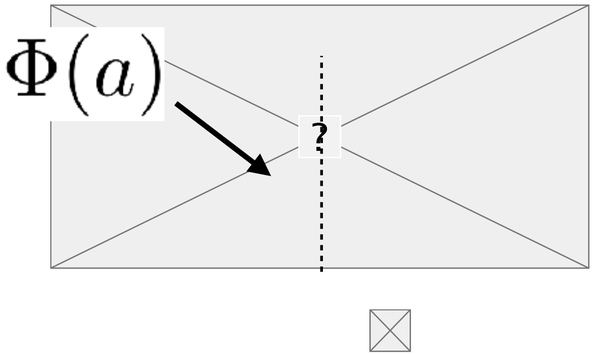
$P(-a \leq Z \leq a) > 0$

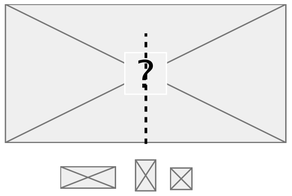
P = ?

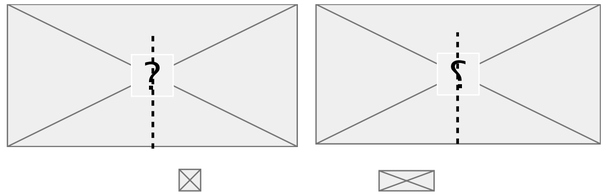
Interval Probability

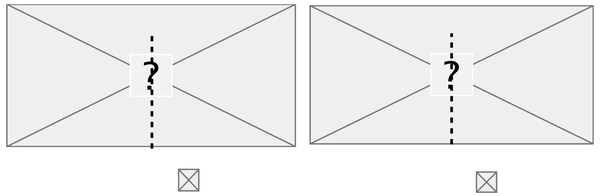
$$Z \sim \mathcal{N}(0,1)$$

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





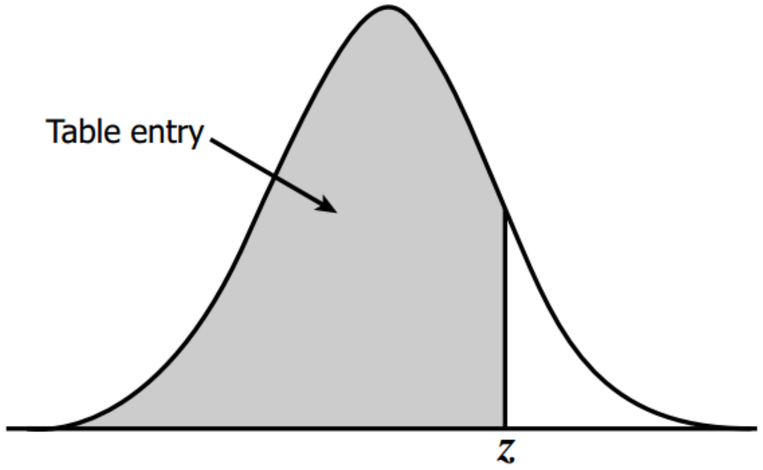




Calculating $\Phi(a)$

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
```

$\Phi(2)$

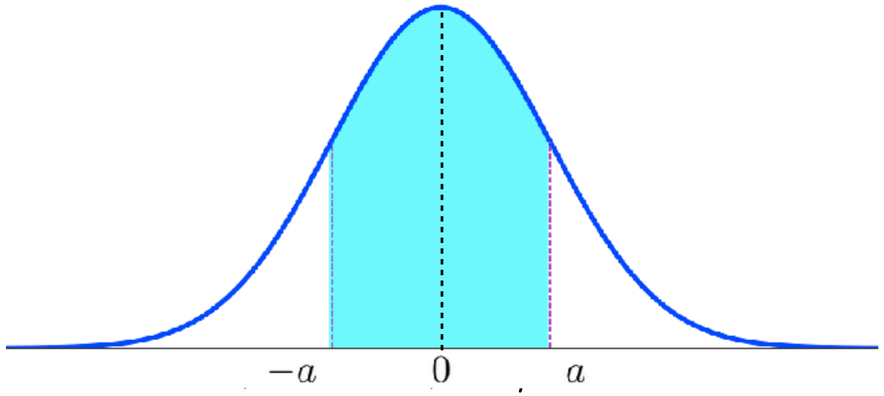
```
norm.cdf(2)
0.977249868052
```

$\Phi(3)$

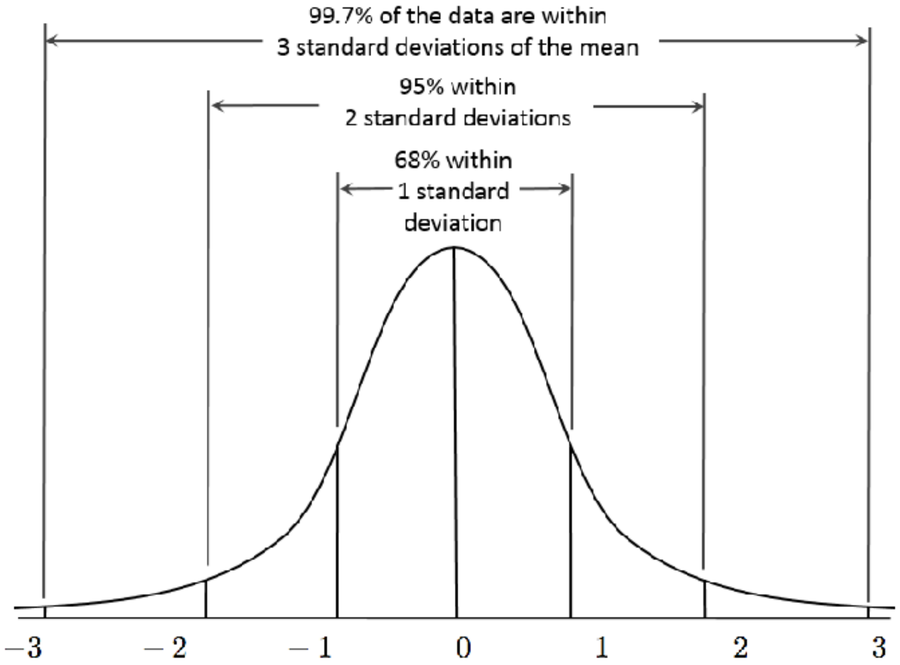
```
norm.cdf(3)
0.998650101968
```

68 - 95 - 99.7 Rule

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



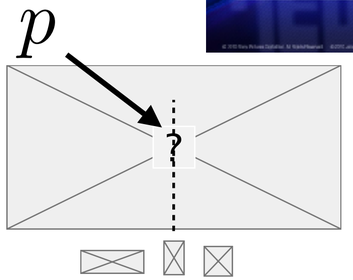
a	P(-a ≤ Z ≤ a)
1	2 · 0.8413 - 1 = 0.682
2	2 · 0.9772 - 1 = 0.9544
3	2 · 0.9987 - 1 = 0.9974



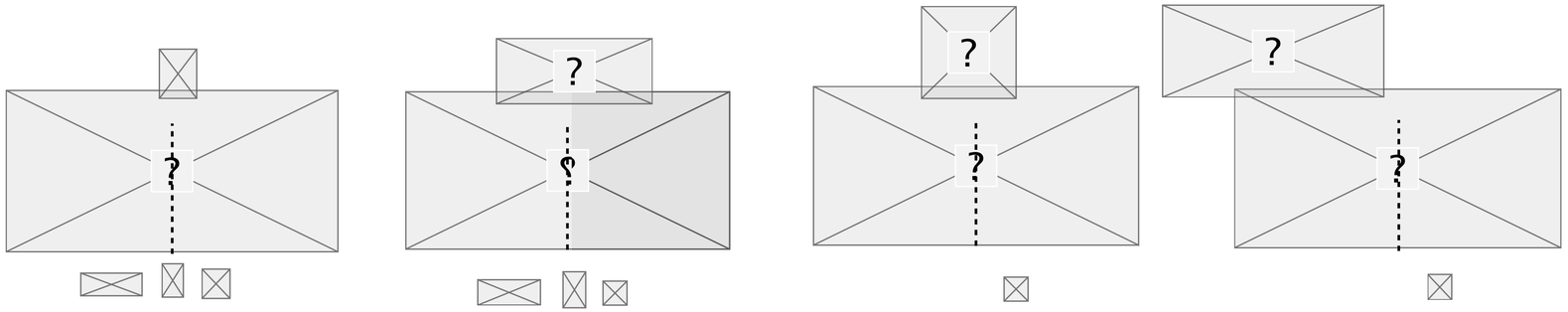
Interval → Probability

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

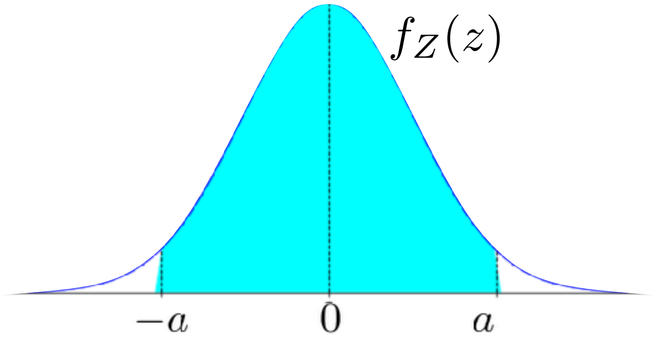
$X \sim \mathcal{N}_{\mu,\sigma^2} \qquad \mathcal{N}(\mu, \sigma^2)$

X, Z

normal

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

Standard Normal



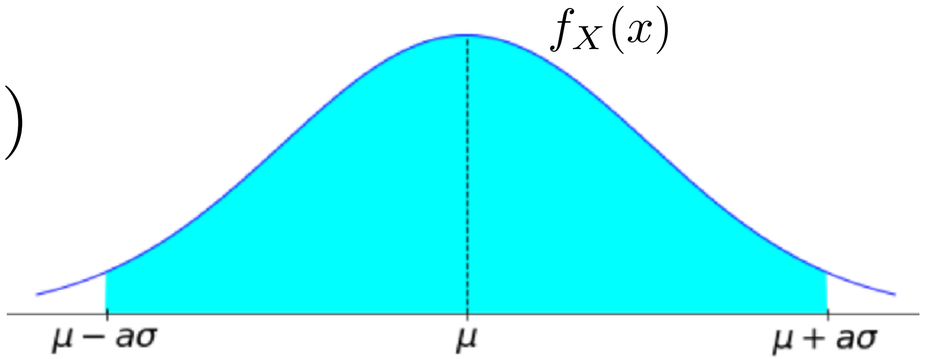
$$P(\mu - a\sigma \leq X \leq \mu + a\sigma) = P(-a\sigma \leq X - \mu \leq a\sigma)$$

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

$\overline{X}$

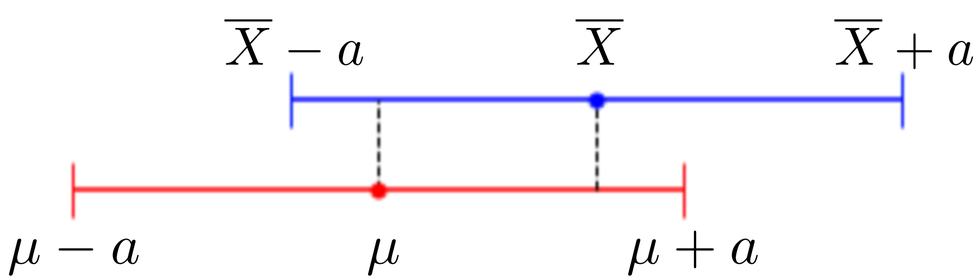
- Roughly normal
- Centered at sample mean $\mu_{\overline{X}} = \mu$
- Standard deviation $V(\overline{X}) = \frac{\sigma^2}{n} \quad \sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}$

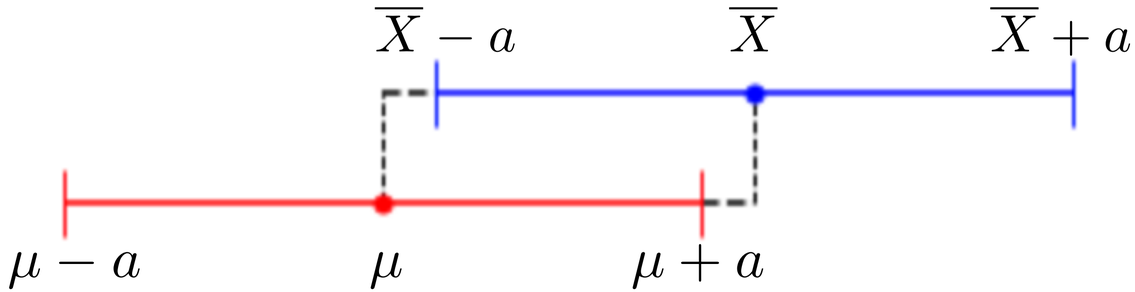
Standard error

Proximity is Reciprocal

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

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

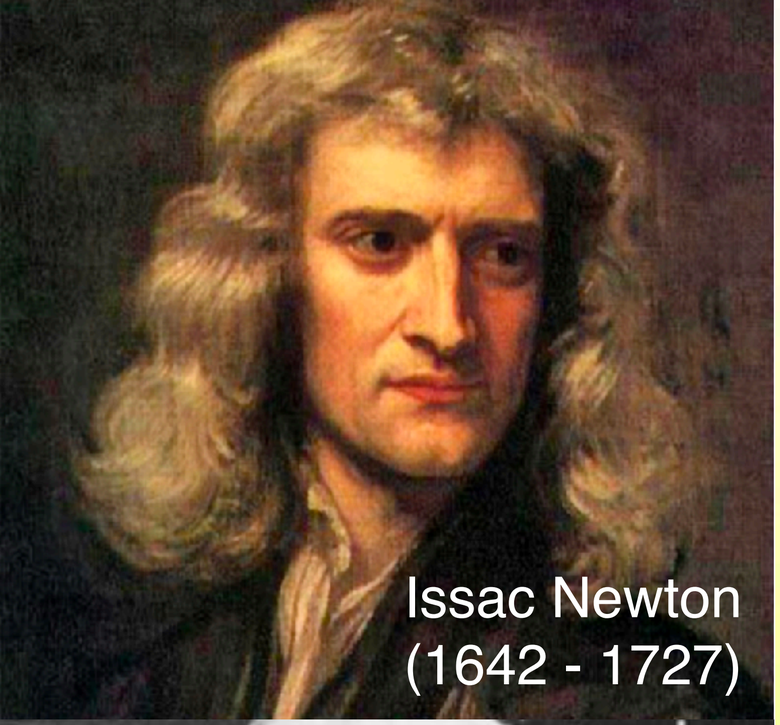




$\overline{X} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$

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

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

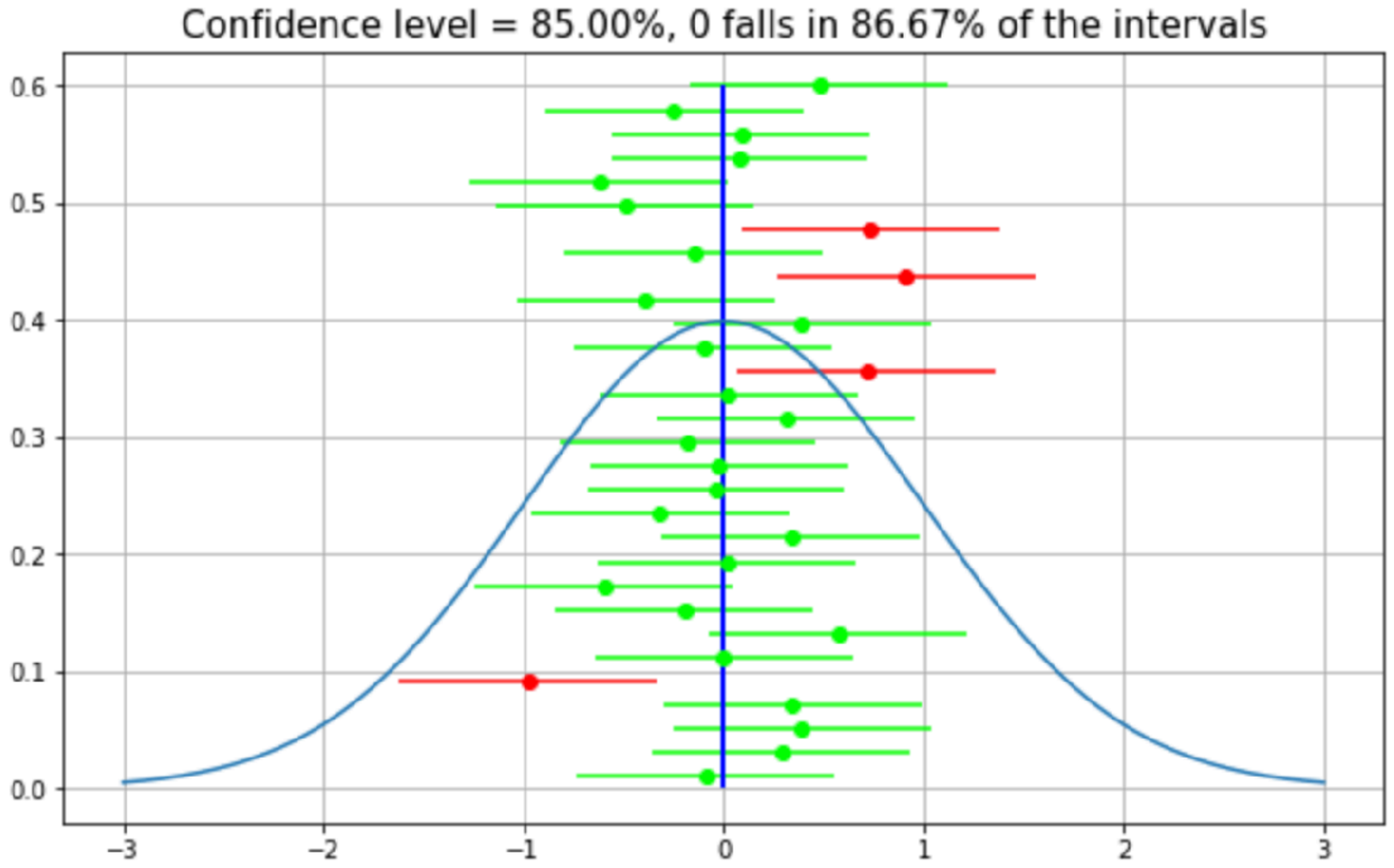
$$\left|\overline{X} - \mu\right| < z_p \sigma_{\overline{X}}$$

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

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