Statistical Data Analysis.
1. Distributions, statistics, and their properties.

Evgeniy Riabenko
riabenko.e@gmail.com

2018

Why taking this course

- concrete statistical methods for different types of problems
- assumptions, limitations, and applicability of different methods
- statistical thinking

Why taking this course

- concrete statistical methods for different types of problems
- assumptions, limitations, and applicability of different methods

  (Marriott, 1974): If the results disagree with informed opinion, do not admit a simple logical interpretation, and do not show up clearly in a graphical presentation, they are probably wrong. There is no magic about numerical methods, and many ways in which they can break down. They are a valuable aid to the interpretation of data, not sausage machines automatically transforming bodies of numbers into packets of scientific fact.
- statistical thinking

Why taking this course

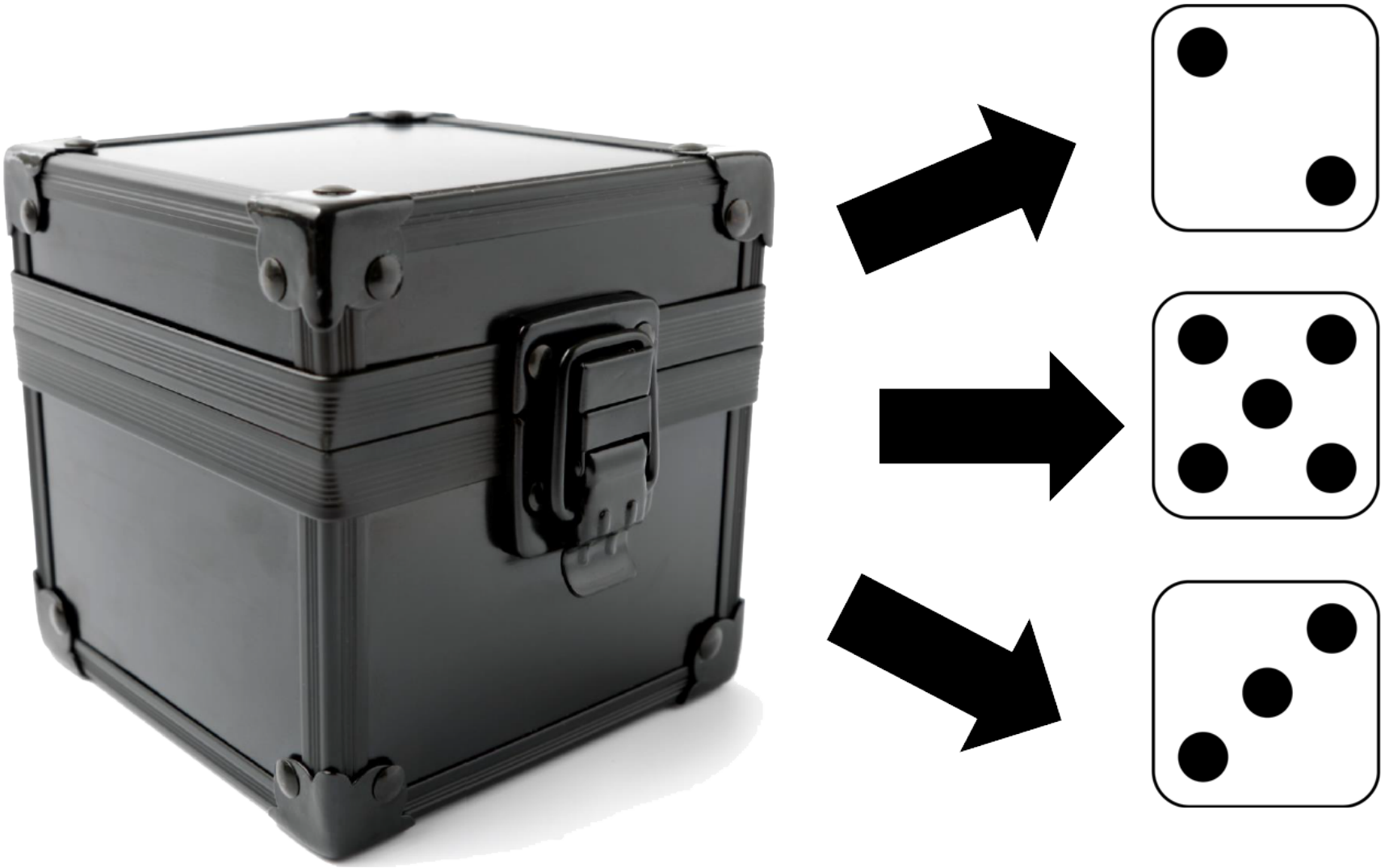
- concrete statistical methods for different types of problems
- assumptions, limitations, and applicability of different methods
- statistical thinking

  (Kahneman et al., 2002): exposure to statistical thinking allows people to rely more on logic than on intuition, helps to correct judgement biases and to find more rational and less stereotypical solutions to problems.

Randomness

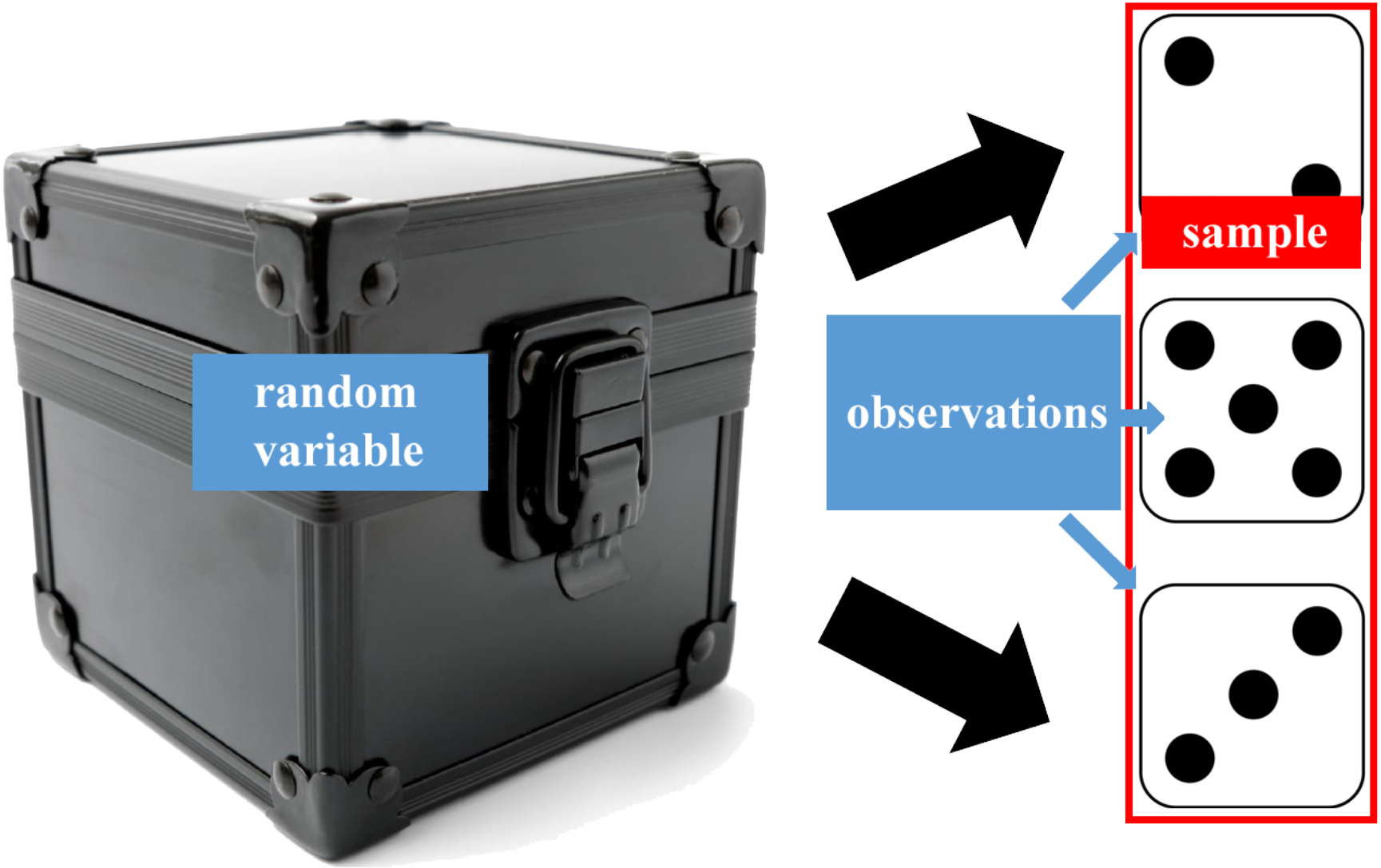
Randomness

Randomness

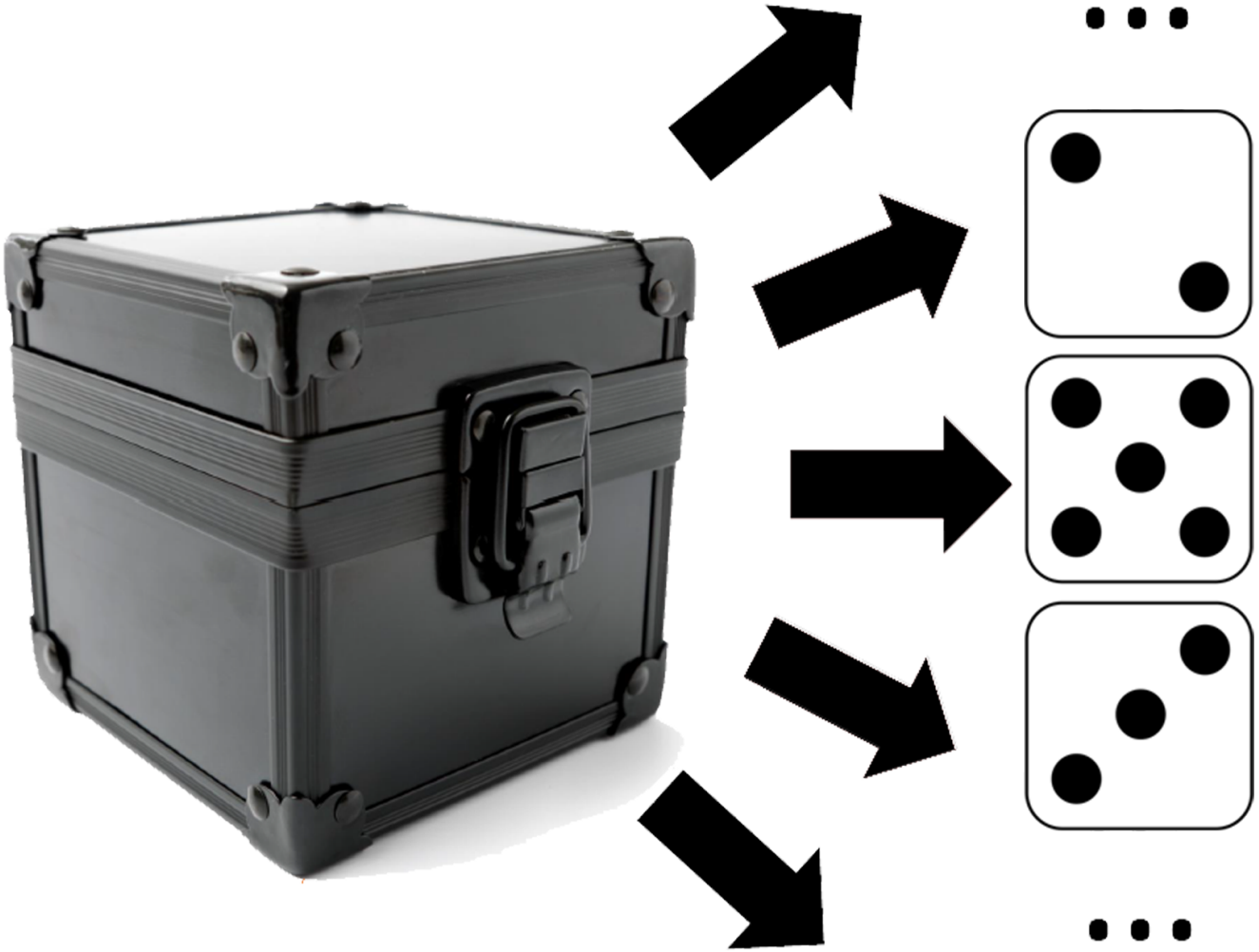


Randomness

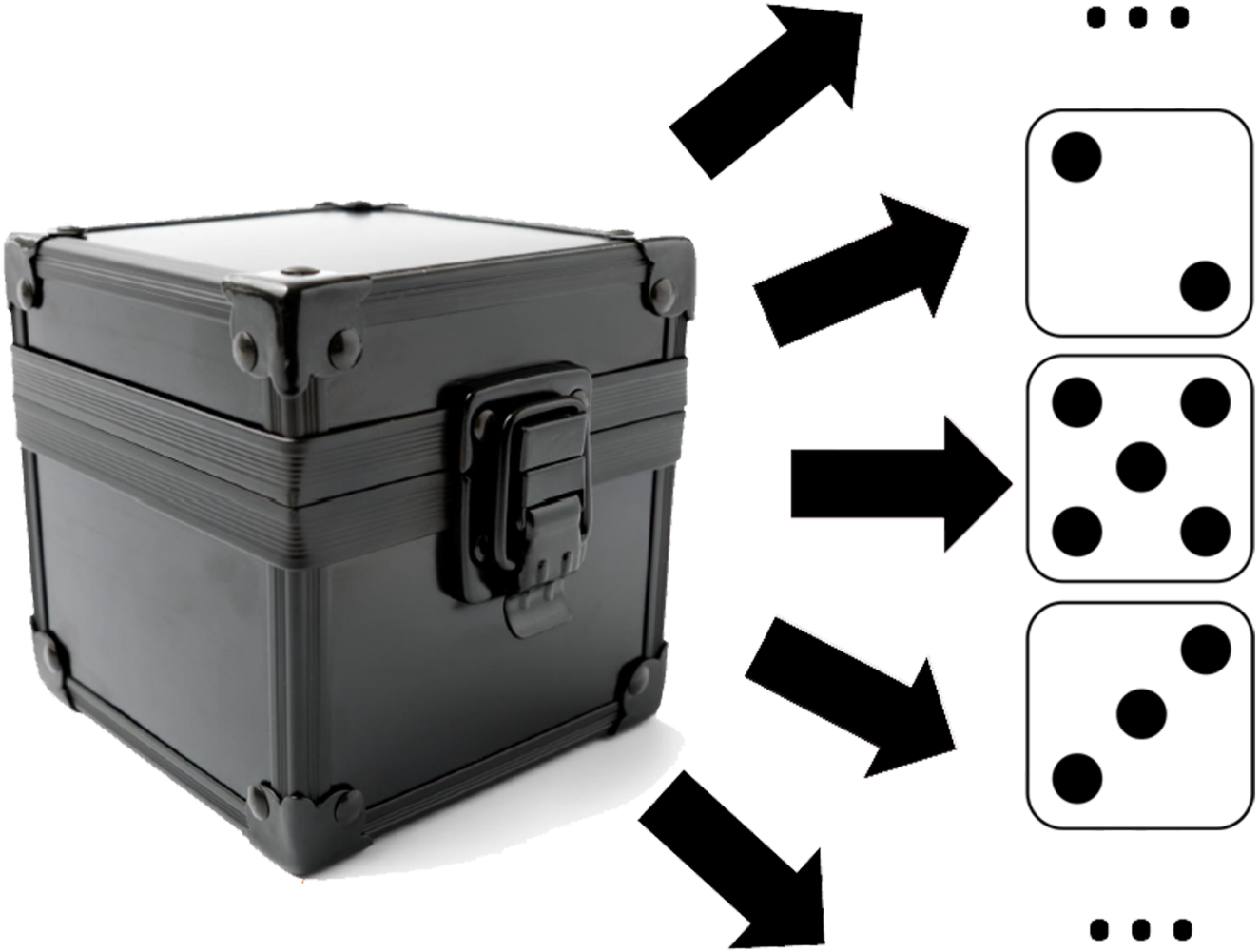
Randomness



Studying randomness

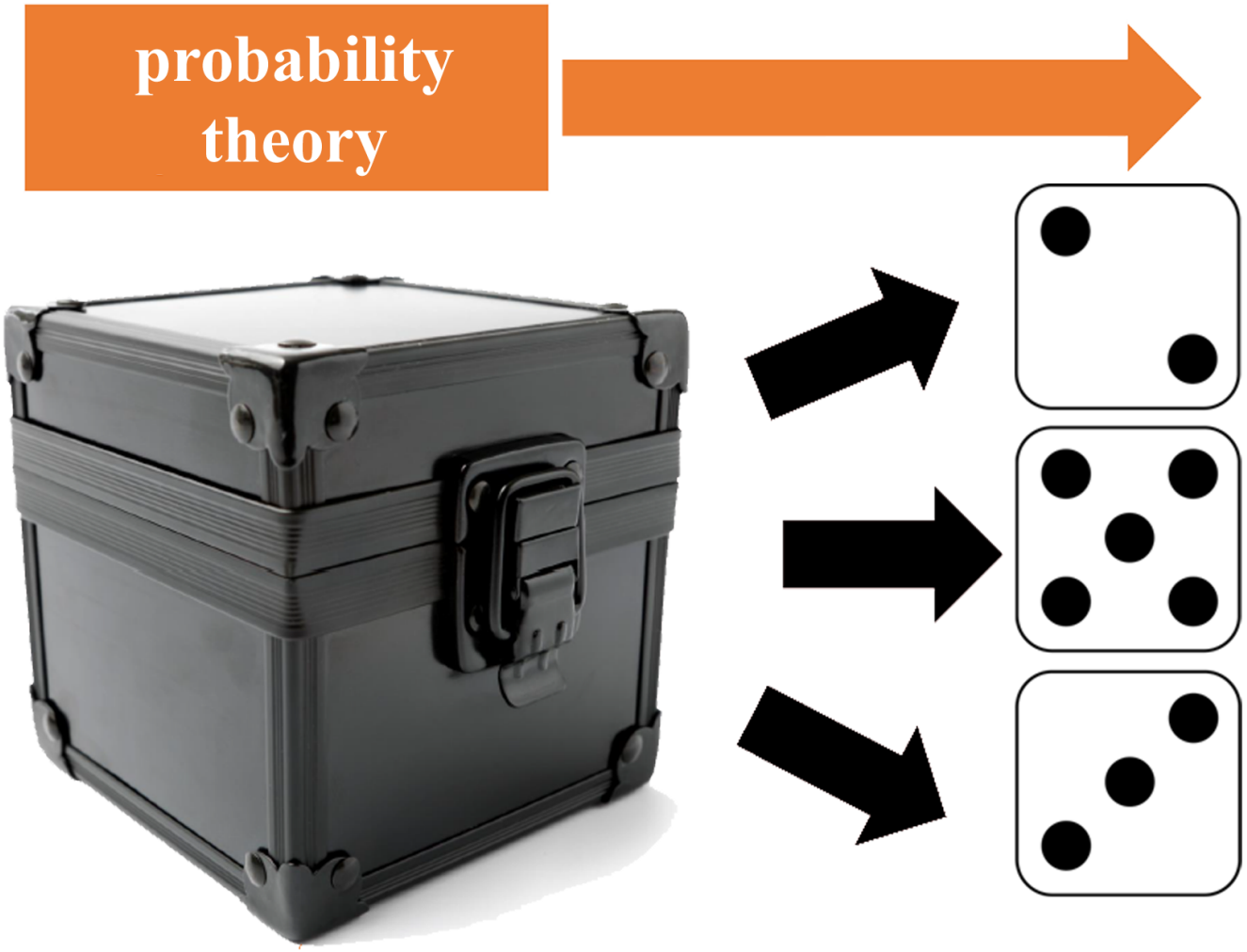


Studying randomness

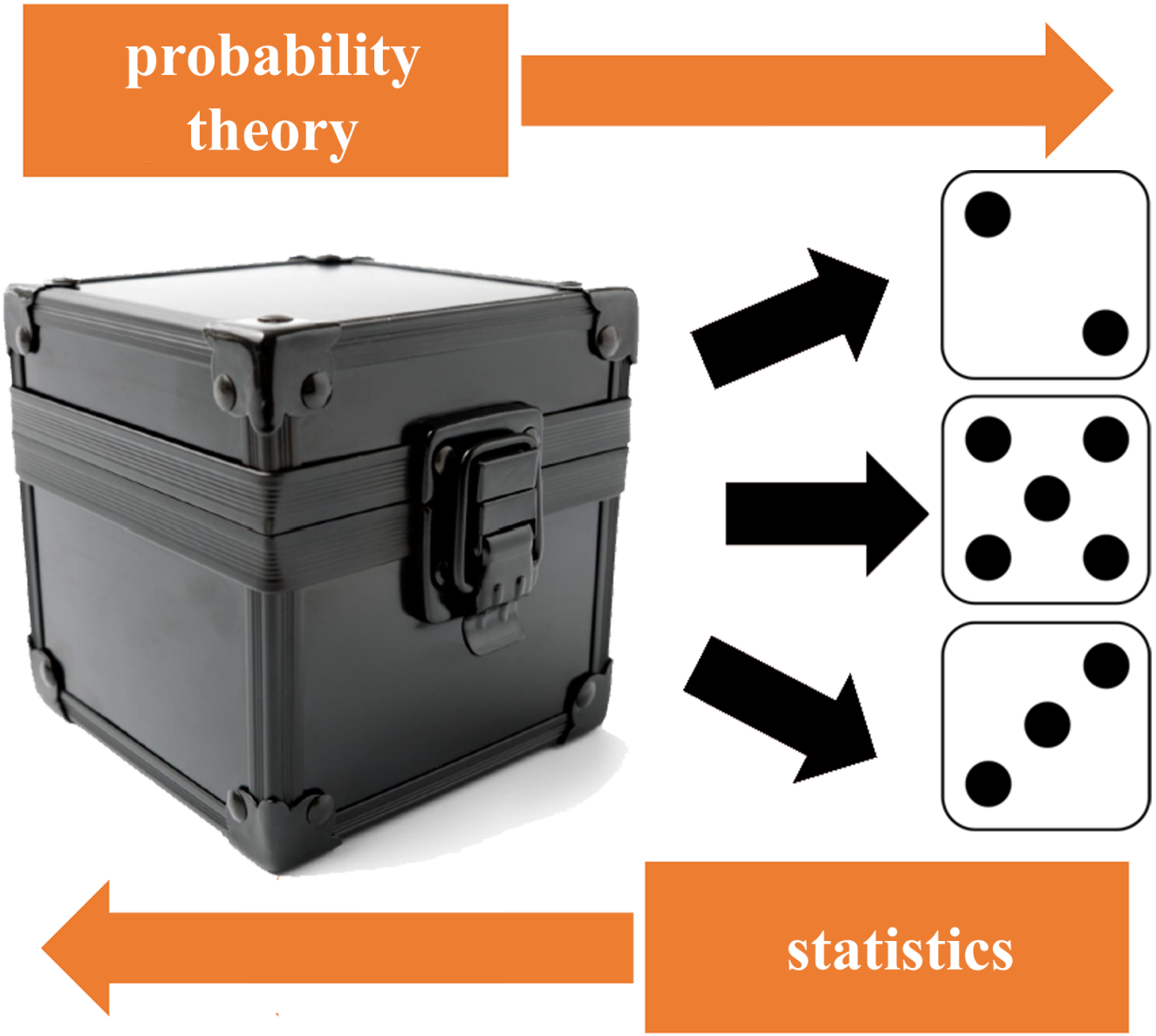


Probability of the event A — a proportion of the outcome A in the infinite sequence of identical trials.

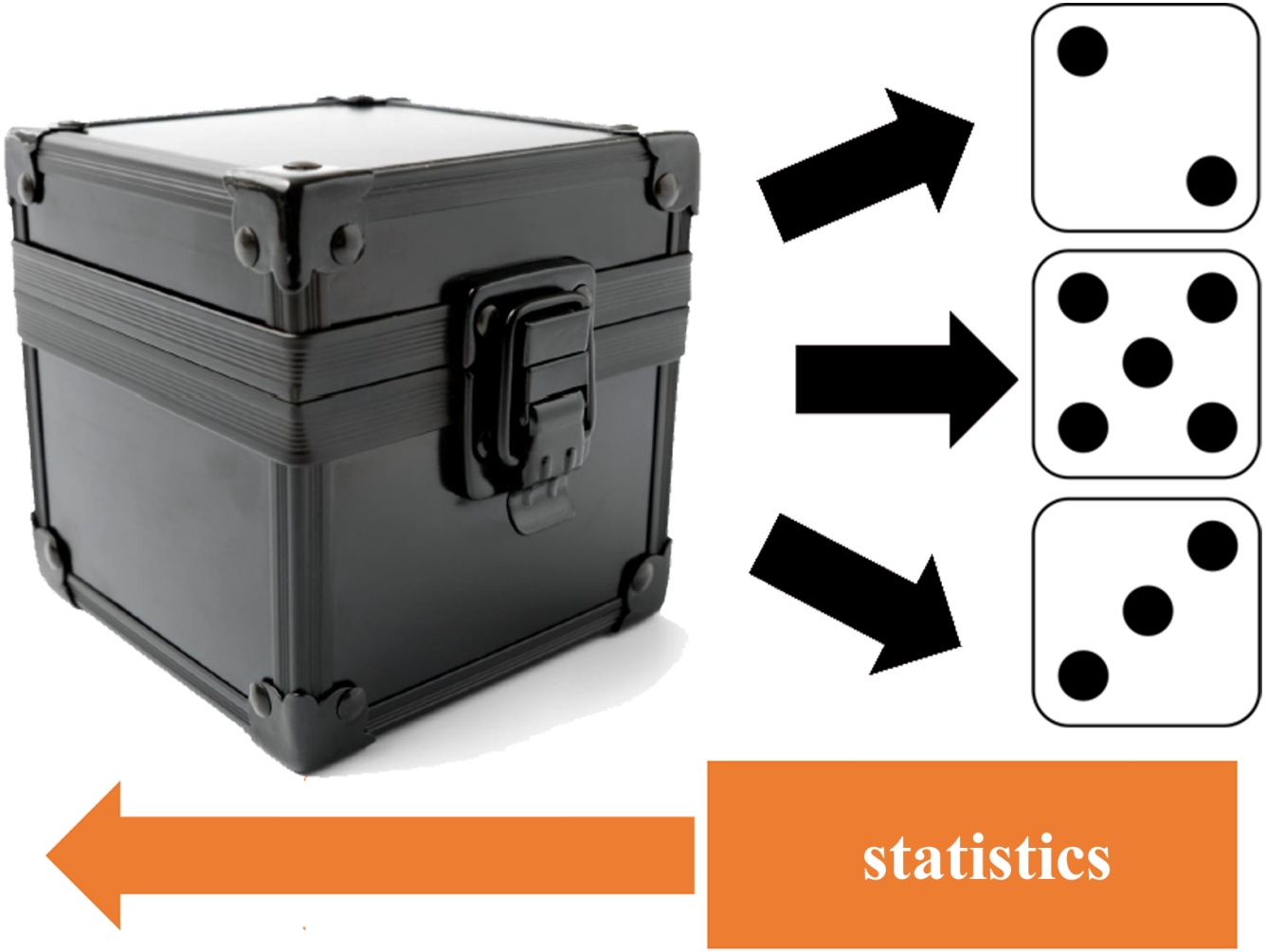
Studying randomness



Studying randomness



Studying randomness



Law of large numbers: for large samples frequencies of events are close to their probabilities.

Describing random variables

Discrete random variable X has a countable set of possible values $A = \{a_1, a_2, \dots\}$ with associated probabilities $p_1, p_2, \dots, \sum_i p_1 = 1$.

$f_X(a_i) = \mathbf{P}(X = a_i) = p_i$ — **probability mass function**.

Continuous random variable is defined with **(cumulative) distribution function**:

$$F_X(x) = \mathbf{P}(X \leqslant x)$$

or **density**:

$$f_X(x) : \int_a^b f_X(x)\,dx = \mathbf{P}(a \leqslant X \leqslant b)$$

.

Characteristics of distributions

- **expected value** — average value of X:

$$\mathbb{E}X = \int x\, dF(x)$$

- **variance** — a measure of spread of X:

$$\mathbb{D}X = \mathbb{E}\left((X - \mathbb{E}X)^2\right)$$

- α **quantile**, $\alpha \in (0, 1)$:

$$X_\alpha\colon \quad \mathbf{P}(X \leqslant X_\alpha) \geqslant \alpha, \;\; \mathbf{P}(X \geqslant X_\alpha) \geqslant 1 - \alpha$$

  equivalent definition:

$$X_\alpha = F^{-1}(\alpha) = \inf\{x\colon F(x) \geqslant \alpha\}$$

- **median** — 0.5 quantile, the middle point of the distribution:

$$\operatorname{med} X\colon \quad \mathbf{P}(X \leqslant \operatorname{med} X) \geqslant 0.5, \;\; \mathbf{P}(X \geqslant \operatorname{med} X) \geqslant 0.5$$

- **interquartile range**:

$$IQR = X_{0.75} - X_{0.25}$$

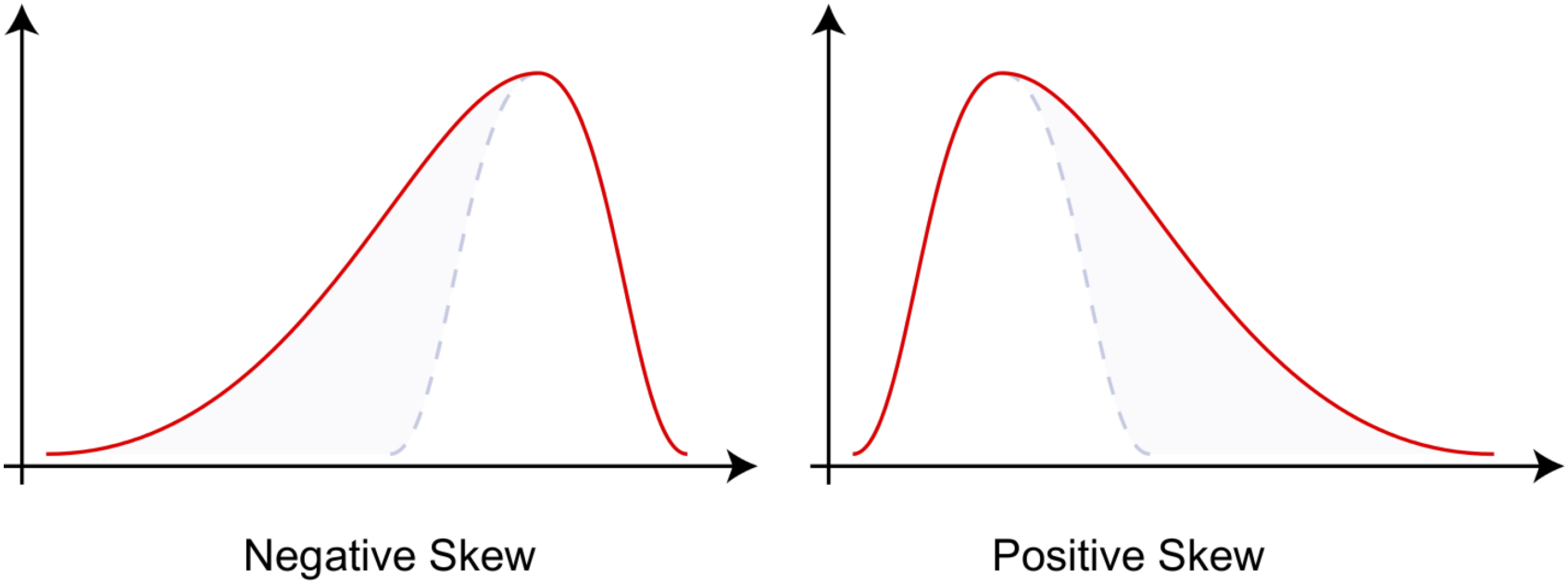
- **mode** — maximum of probability mass function or density:

$$\operatorname{mode} X = \operatorname*{argmax}_x f_X(x)$$

Characteristics of distributions

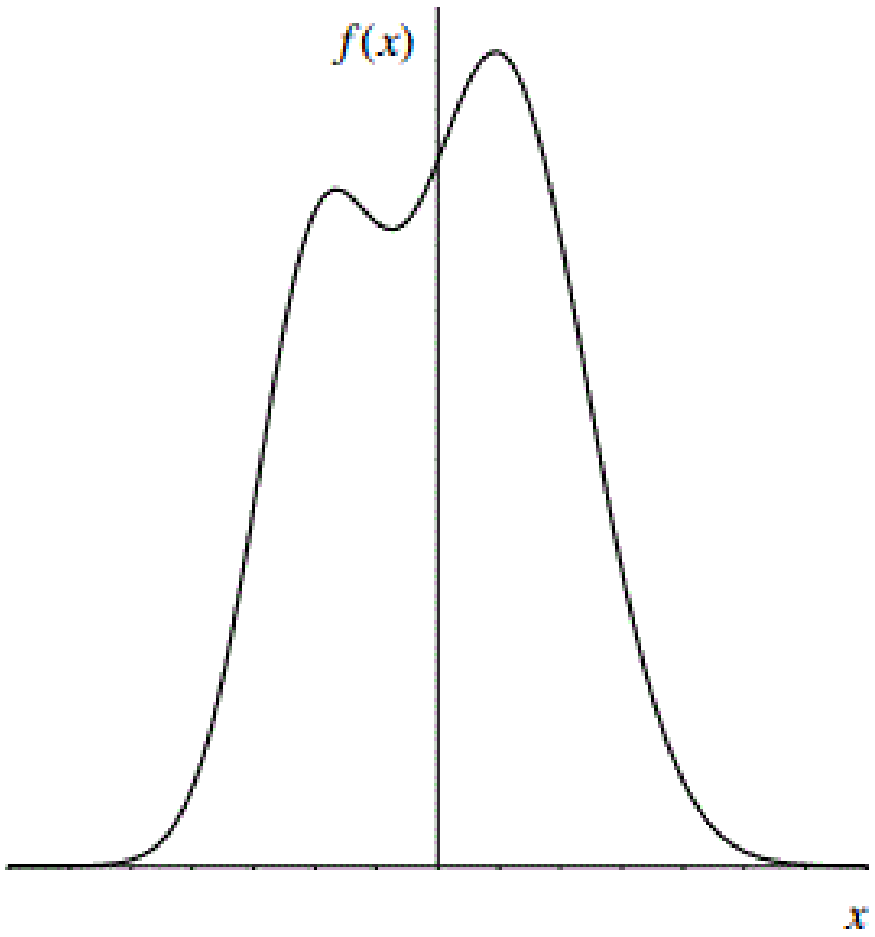
- **skewness**:

$$\gamma_1 = \mathbb{E}\left(\frac{X - \mathbb{E}X}{\sqrt{\mathbb{D}X}}\right)^3$$



Characteristics of distributions

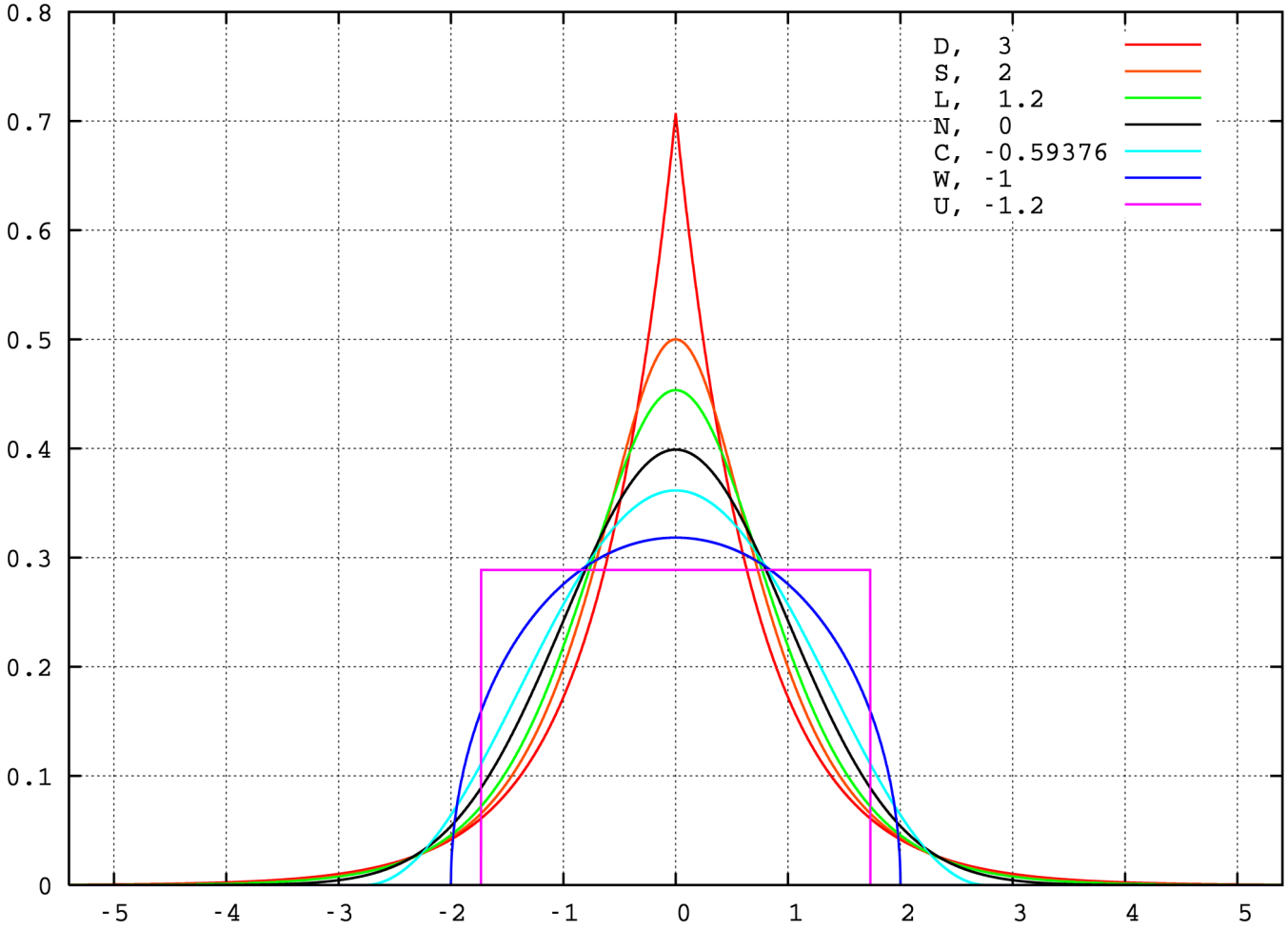
$\gamma_1 = 0$ — necessary, but not sufficient for symmetry:



Characteristics of distributions

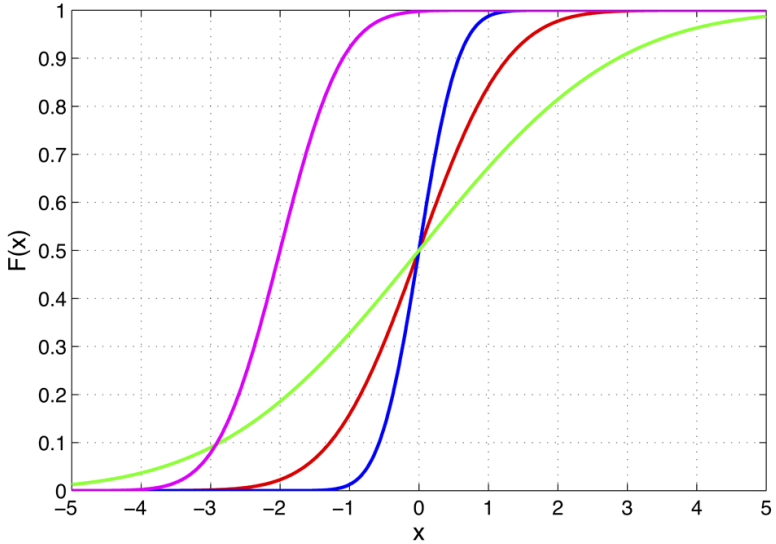
- **excess** (without -3 — kurtosis):

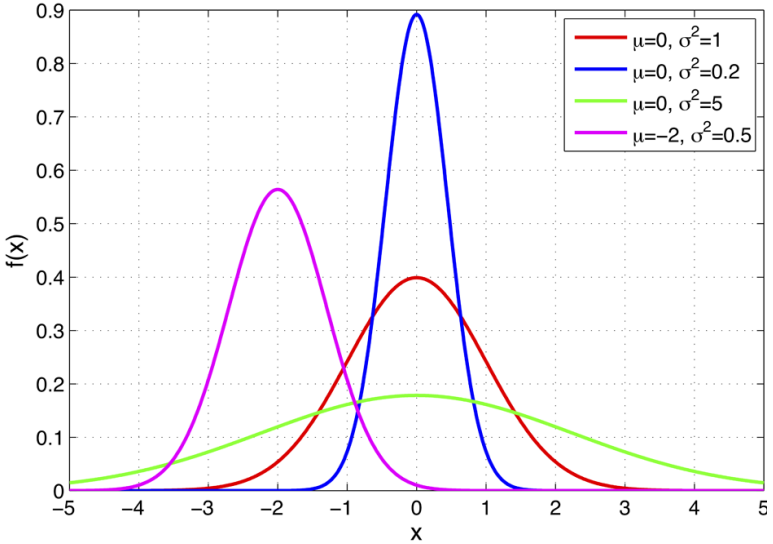
$$\gamma_2 = \frac{\mathbb{E}\left(X - \mathbb{E}X\right)^4}{\left(\mathbb{D}X\right)^2} - 3$$



Gaussian distribution

$$X \in \mathbb{R} \sim N\left(\mu, \sigma^2\right), \ \sigma^2 > 0$$





$$F(x) = \Phi\left(\frac{x-\mu}{\sigma}\right) \qquad \Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}} dt$$

$$f(x) = \frac{1}{\sigma}\phi\left(\frac{x-\mu}{\sigma}\right) \qquad \phi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

Gaussian distribution

- a limiting distribution of a sum of weakly dependent random variables
- $\mathbb{E}X = \operatorname{med} X = \operatorname{mode} X = \mu$, $\mathbb{D}X = \sigma^2$, all the higher moments are zero
- if $X_1, \ldots, X_n$ are independent, $X_i \sim N\left(\mu_i, \sigma_i^2\right)$, then $\forall a_1, \ldots, a_n$

$$\sum_{i=1}^{n} a_i X_i \sim N\left(\sum_{i=1}^{n} a_i \mu_i, \sum_{i=1}^{n} a_i^2 \sigma_i^2\right)$$

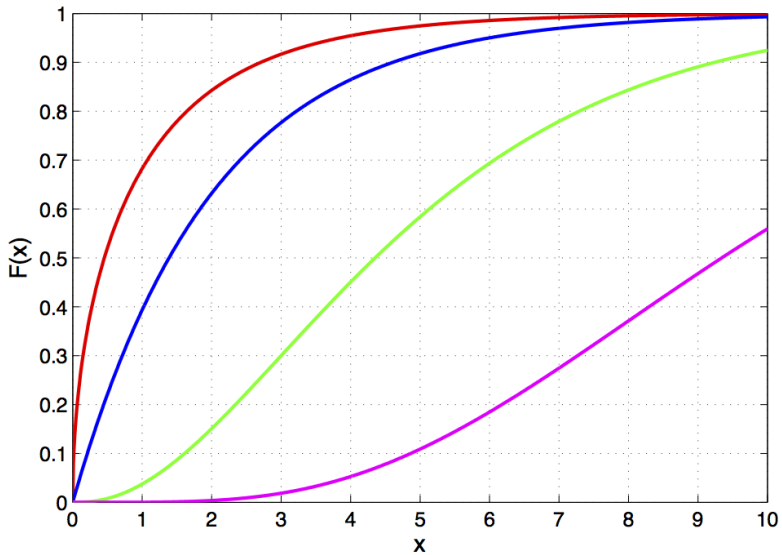
- central limit theorem: if $X_1, \ldots, X_n$ are i.i.d. with $\mathbb{E}X$ and $\mathbb{D}X < \infty$, then

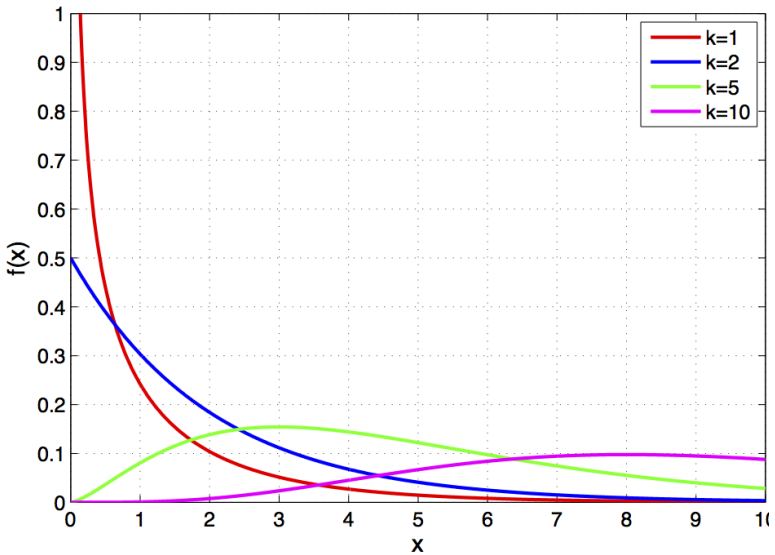
$$\frac{1}{n}\sum_{i=1}^{n} X_i \sim\approx N\left(\mathbb{E}X, \frac{\mathbb{D}X}{n}\right)$$

- example: measurement error

Chi-squared distribution

$X \in \mathbb{R}_+ \sim \chi^2_k,\ k \in \mathbb{N}$





$$F(x) = \frac{1}{\Gamma\left(\frac{k}{2}\right)}\gamma\left(\frac{k}{2}, \frac{x}{2}\right)$$

$$f(x) = \frac{1}{2^{\frac{k}{2}}\Gamma\left(\frac{k}{2}\right)}x^{\frac{k}{2}-1}e^{-\frac{x}{2}}$$

$\Gamma(x) = \int_0^\infty t^{x-1}e^{-t}dt$ — gamma function

$\gamma(a, x) = \int_0^x e^{-t}t^{a-1}dt$ — lower incomplete gamma function

Chi-squared distribution

- if $X_1, \ldots, X_k$ are i.i.d., $X_i \sim N(0,1)$, then

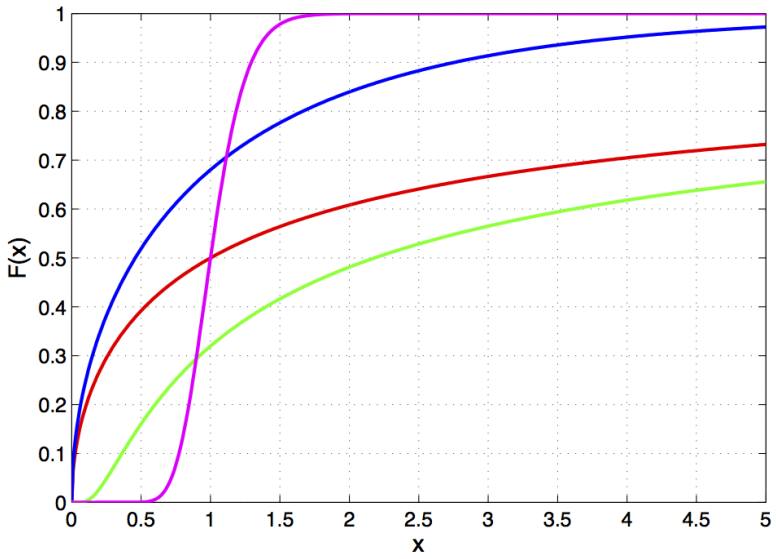
$$\sum_{i=1}^{k} X_i^2 \sim \chi_k^2$$

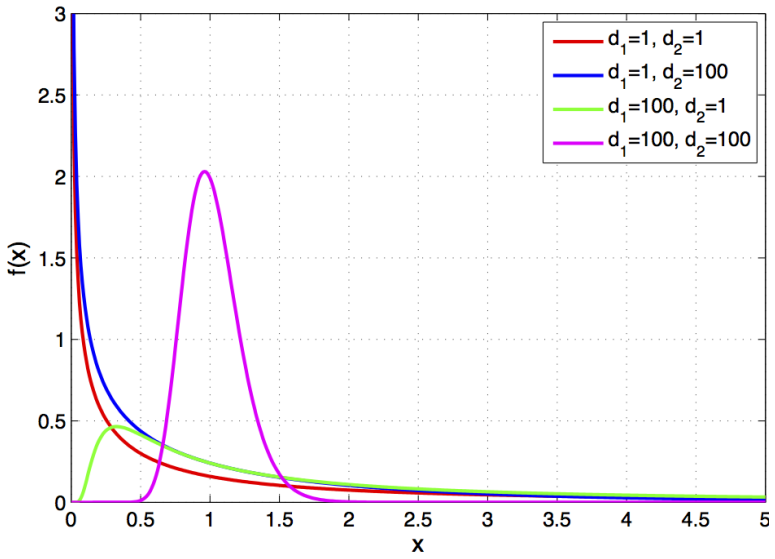
- example: normalized sample variance:

$$(n-1)\frac{S_n^2}{\sigma^2} \sim \chi_{n-1}^2$$

F-distribution

$$X \in \mathbb{R}_+ \sim F(d_1, d_2), \; d_1, d_2 > 0$$





$$F(x) = I_{\frac{d_1 x}{d_1 x + d_2}}\left(\frac{d_1}{2}, \frac{d_2}{2}\right)$$

$$f(x) = \sqrt{\frac{(d_1 x)^{d_1} d_2^{d_2}}{(d_1 x + d_2)^{d_1 + d_2}}} \Big/ x B\left(\frac{d_1}{2}, \frac{d_2}{2}\right)$$

$B(a,b) = \int_0^1 t^{a-1}(1-t)^{b-1}\,dt$ — beta function

$I_x(a,b) = \frac{B(x;a,b)}{B(a,b)}$ — regularized incomplete beta function

$B(x;a,b) = \int_0^x t^{a-1}(1-t)^{b-1}\,dt$ — incomplete beta function

F-distribution

- if $X_1 \sim \chi^2_{d_1}, \;\; X_2 \sim \chi^2_{d_2}, \quad X_1$ and X_2 are independent, then

$$\frac{X_1/d_1}{X_2/d_2} \sim F(d_1, d_2)$$

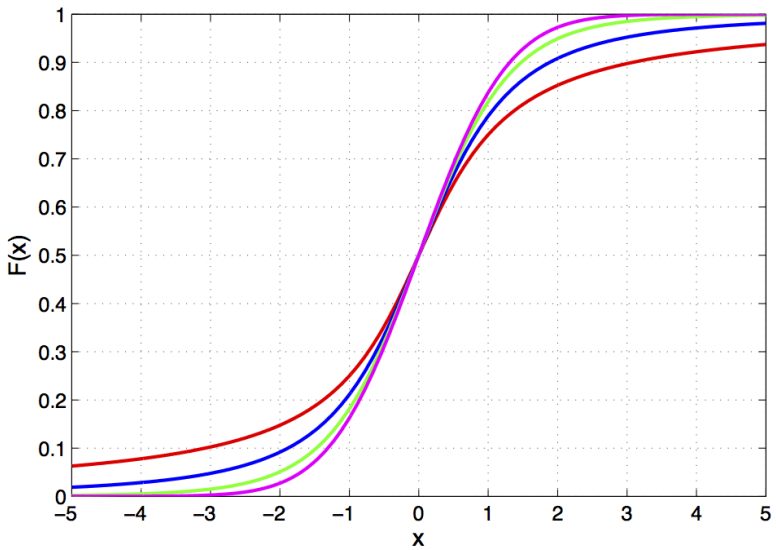
- if $X \sim F(d_1, d_2)$, then

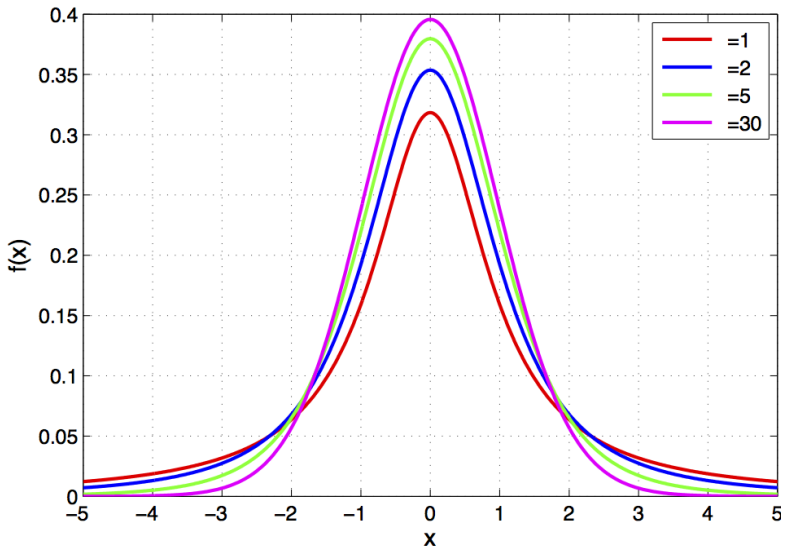
$$Y = \lim_{d_2 \to \infty} d_1 X \sim \chi^2_{d_1}$$

- $F(x, d_1, d_2) = F(1/x, d_2, d_1)$
- arises frequently in ANOVA and regression tests

Student's t-distribution

$$X \in \mathbb{R} \sim St\left(\nu\right),\ \nu > 0$$





$$F\left(x\right) = \frac{1}{2} + x\Gamma\left(\frac{\nu+1}{2}\right)$$

$$f\left(x\right) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\Gamma\left(\frac{\nu}{2}\right)}\left(1+\frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}$$

Student's t-distribution

- $\mathbb{E}X = 0$ for $\nu > 1$, $\operatorname{med} X = \operatorname{mode} X = 0$ always
- if $Z \sim N(0,1)$ are $V \sim \chi^2_\nu$ independent, then

$$T = \frac{Z}{\sqrt{V/\nu}} \sim St(\nu)$$

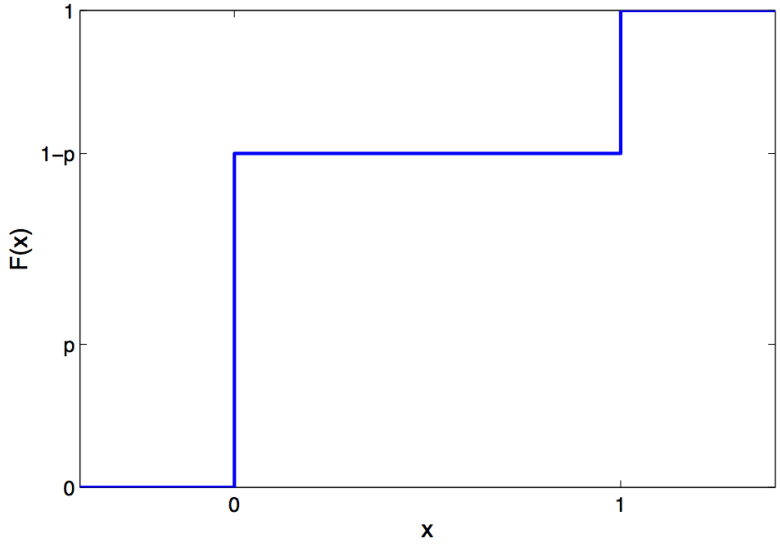
- if $X \sim St(\nu)$, then

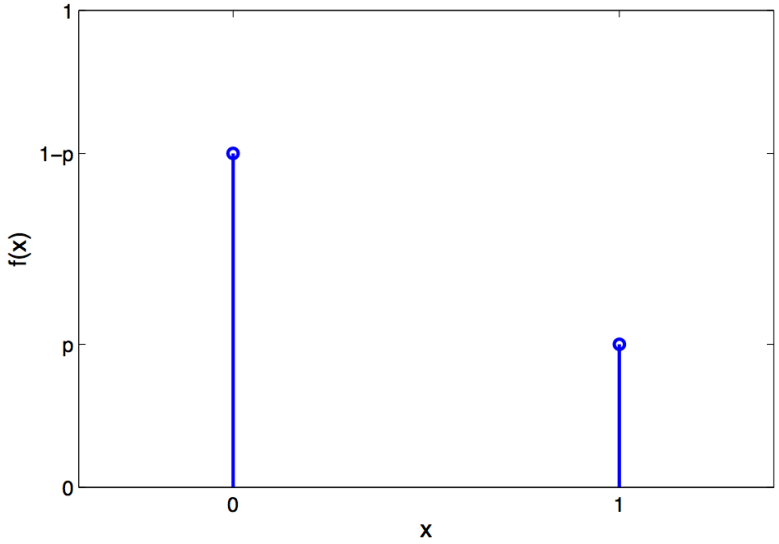
$$Y = \lim_{\nu \to \infty} X \sim N(0,1)$$

- very often used in testing hypotheses about means

Bernoulli distribution

$$X \in \{0,1\} \sim Ber\,(p)\,, \quad p \in (0,1)$$





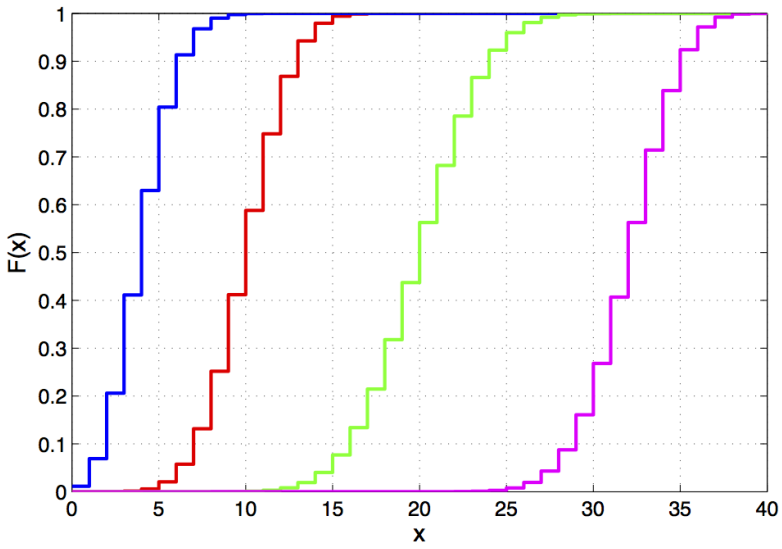
$$F\,(x) = \begin{cases} 0, & x < 0, \\ 1-p, & 0 \leqslant x < 1, \\ 1, & x \geqslant 1. \end{cases}$$

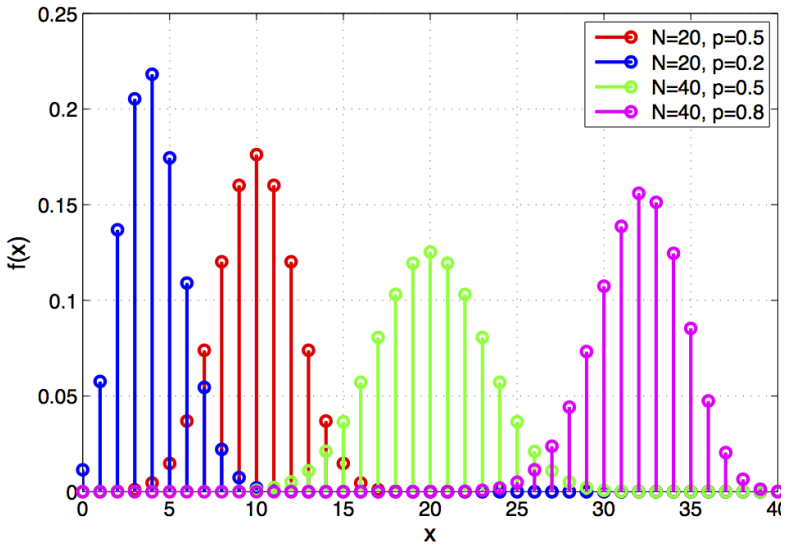
$$f\,(x) = \begin{cases} 1-p, & x = 0, \\ p, & x = 1. \end{cases}$$

- example: coin tossing

Binomial distribution

$$X \in \{0, \ldots, N\} \sim Bin\left(N, p\right), \;\; N \in \mathbb{N}, \;\; p \in [0, 1]$$





$$F\left(x\right) = I_{1-p}\left(N - x, 1 + x\right)$$

$$f\left(x\right) = C_N^x p^x \left(1 - p\right)^{N-x}$$

Binomial distribution

- if $X_1, \ldots, X_n$ are independent, $X_i \sim Ber(p)$, then

$$\sum_{i=1}^{n} X_i \sim Bin(n, p).$$

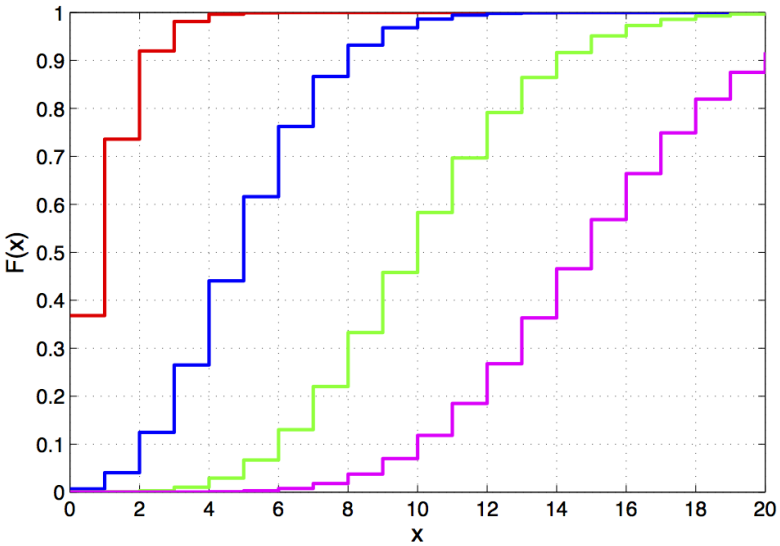
- $Bin\,(1, p) = Ber\,(p)$
- if $N > 20$ and p not too close to 0 or 1, then $X \sim Bin\,(N, p)$ could be well approximated by gaussian:

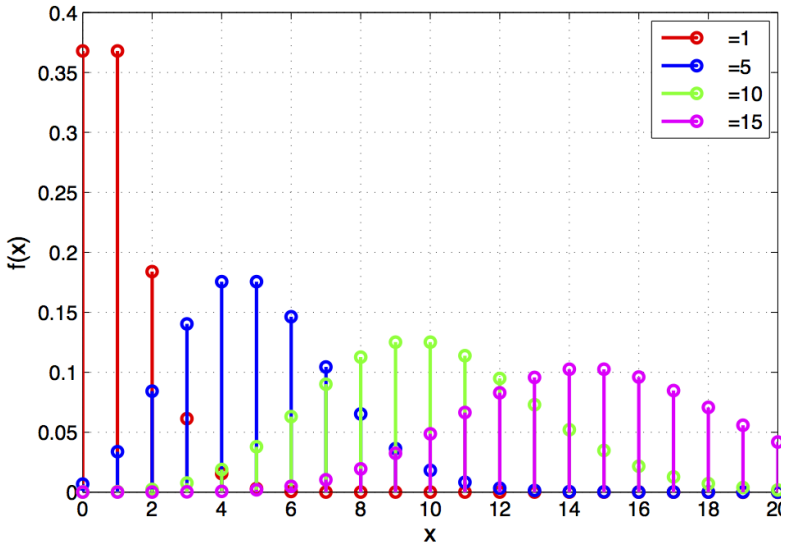
$$F_X\,(x) \approx \Phi\left(\frac{x - Np}{\sqrt{Np\,(1-p)}}\right)$$

- example: number of baskets made in N attempts

Poisson distribution

$$X \in \{0, 1, 2, \dots\} \sim Pois\,(\lambda)\,,\ \lambda > 0$$





$$F\,(x) = e^{-\lambda} \sum_{i=0}^{\lfloor x \rfloor} \frac{\lambda^i}{i!}$$

$$f\,(x) = e^{-\lambda} \frac{\lambda^x}{x!}$$

Poisson distribution

- describes a number of independent events over a fixed temporal or spatial interval
- $\mathbb{E}X = \mathbb{D}X = \lambda$
- if $X_1, \ldots, X_n$ are independent, $X_i \sim Pois\left(\lambda_i\right)$, then

$$\sum_{i=1}^{n} X_i \sim Pois\left(\sum_{i=1}^{n} \lambda_i\right)$$

- if $X \sim Pois\left(\lambda\right)$, $Y = \sqrt{X}$, then for large λ

$$F_Y\left(x\right) \approx \Phi\left(\frac{x - \sqrt{\lambda}}{\sqrt{\lambda}}\right)$$

- example: number of raisins in a bun

Sample

Population — a set of similar items or events which is of interest for some question or experiment.

Sample — finite set of objects selected from the population for measurements.

$$X^n = (X_1, \ldots, X_n)\,.$$

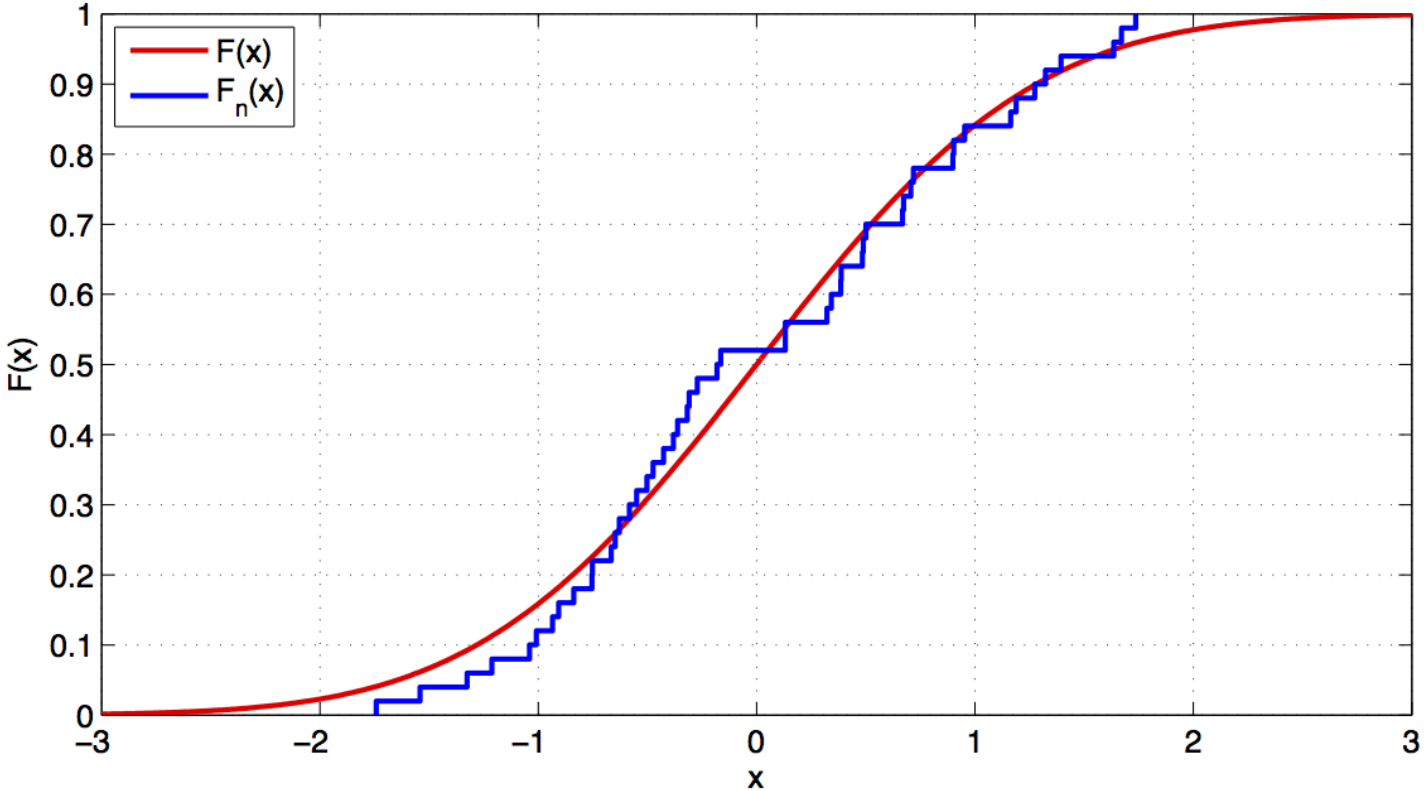
n — **sample size**.

X^n — **simple random sample**, if $X_1, \ldots, X_n$ — independent identically distributed (i.i.d.) random variables.

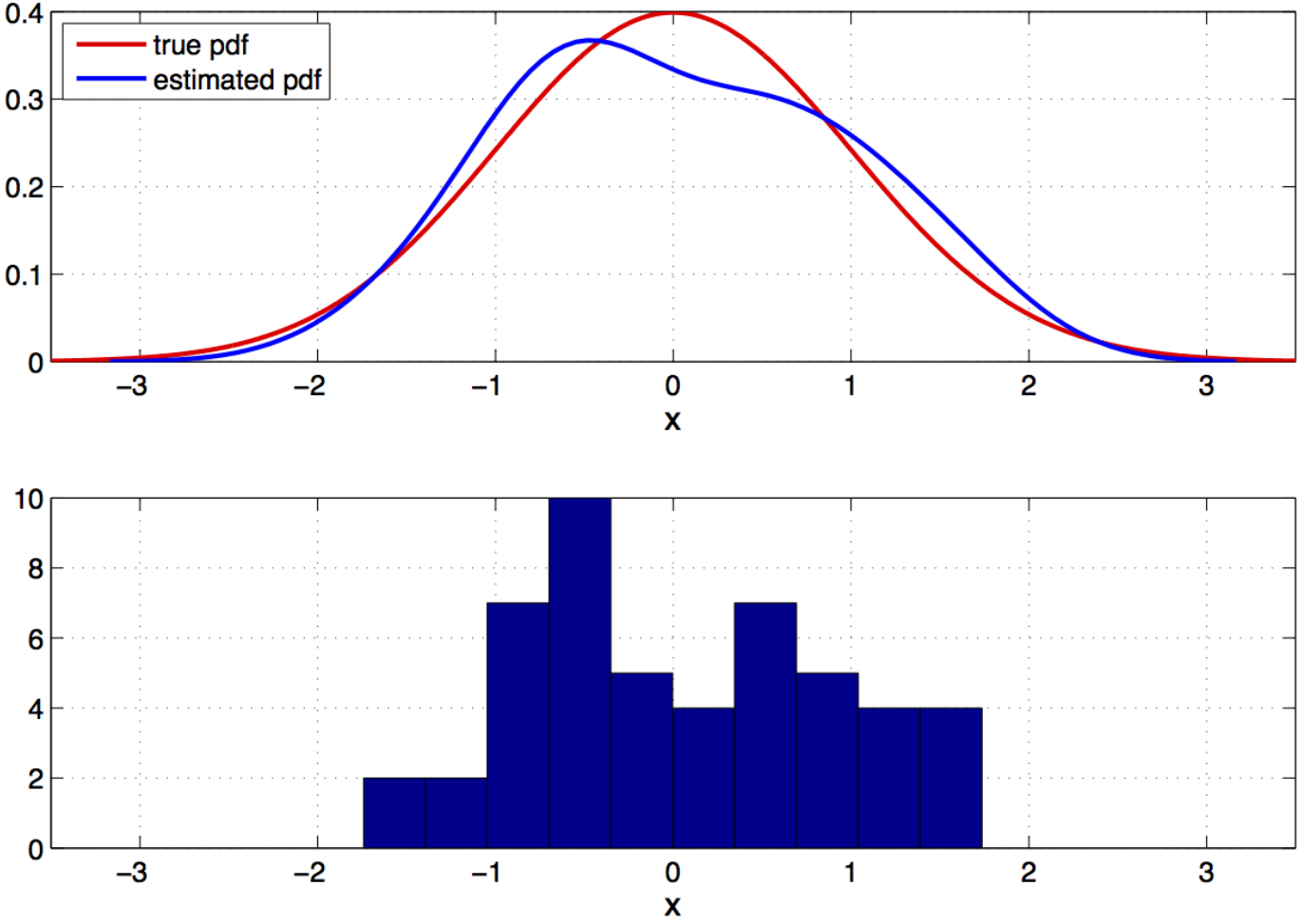
The main goal of statistics is to infer $F_X(x)$ from the observations of the sample.

Distribution function

$F_n(x) = \frac{1}{n}\sum_{i=1}^{n}[X_i \leqslant x]$ — **empirical cumulative distribution function**.



Density


true pdf
estimated pdf
0.4
0.3
0.2
0.1
0
−3
−2
−1
0
1
2
3
x
10
8
6
4
2
0
−3
−2
−1
0
1
2
3
x

Statistic

Statistic $T(X^n)$ — any measurable function of the sample.

- sample mean:

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

- sample variance:

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

sorted sample:

$$X_{(1)} \leqslant X_{(2)} \leqslant \ldots \leqslant X_{(n)}$$

rank of the observation X_i:

$$\operatorname{rank}(X_i) = r\colon X_i = X_{(r)}$$

- kth order statistic: $X_{(k)}$
- sample α-quantile: $X_{([n\alpha])}$
- sample median:

$$m = \begin{cases} X_{(k+1)}, & n = 2k+1, \\ \frac{X_{(k)}+X_{(k+1)}}{2}, & n = 2k. \end{cases}$$

Statistic

- sample interquartile range:

$$IQR_n = X_{([0.75n])} - X_{([0.25n])}$$

- sample skewness:

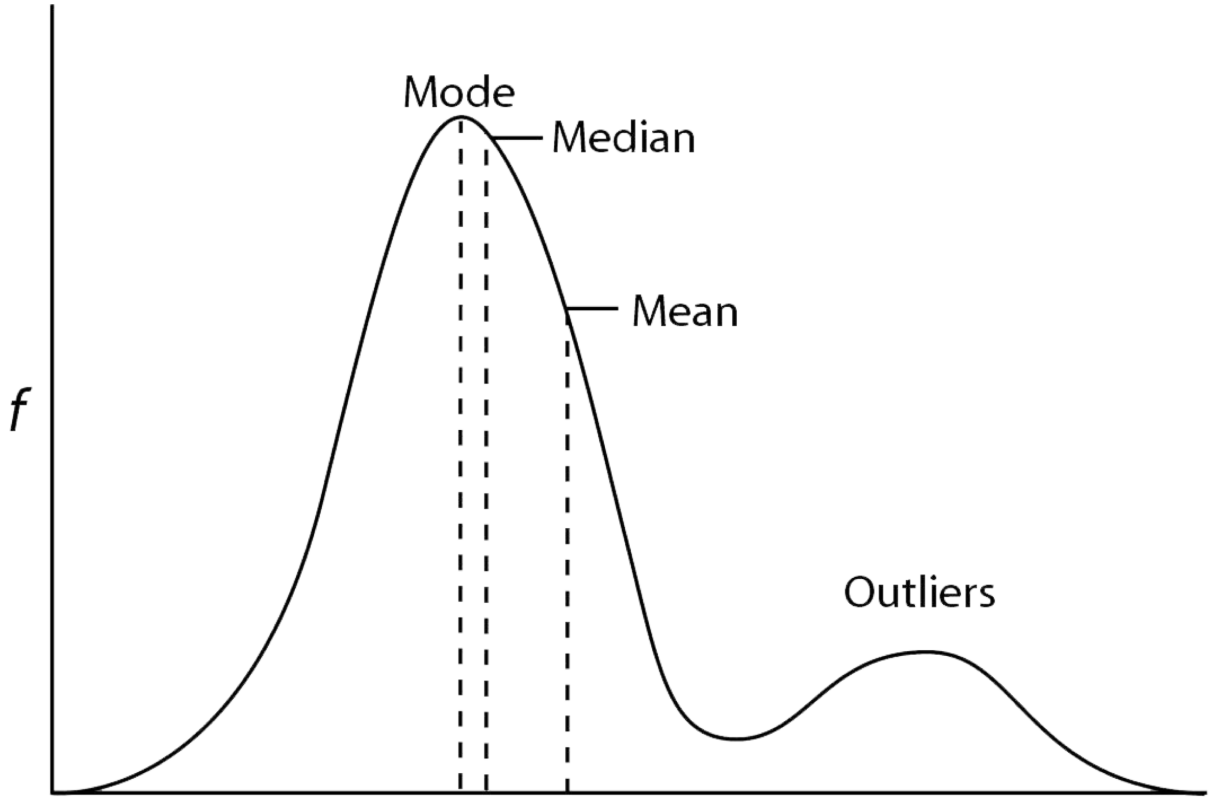
$$g_1 = \frac{\sqrt{n}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^3}{\left(\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2\right)^{3/2}}$$

- sample excess:

$$g_2 = \frac{n\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^4}{\left(\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2\right)^{2}} - 3$$

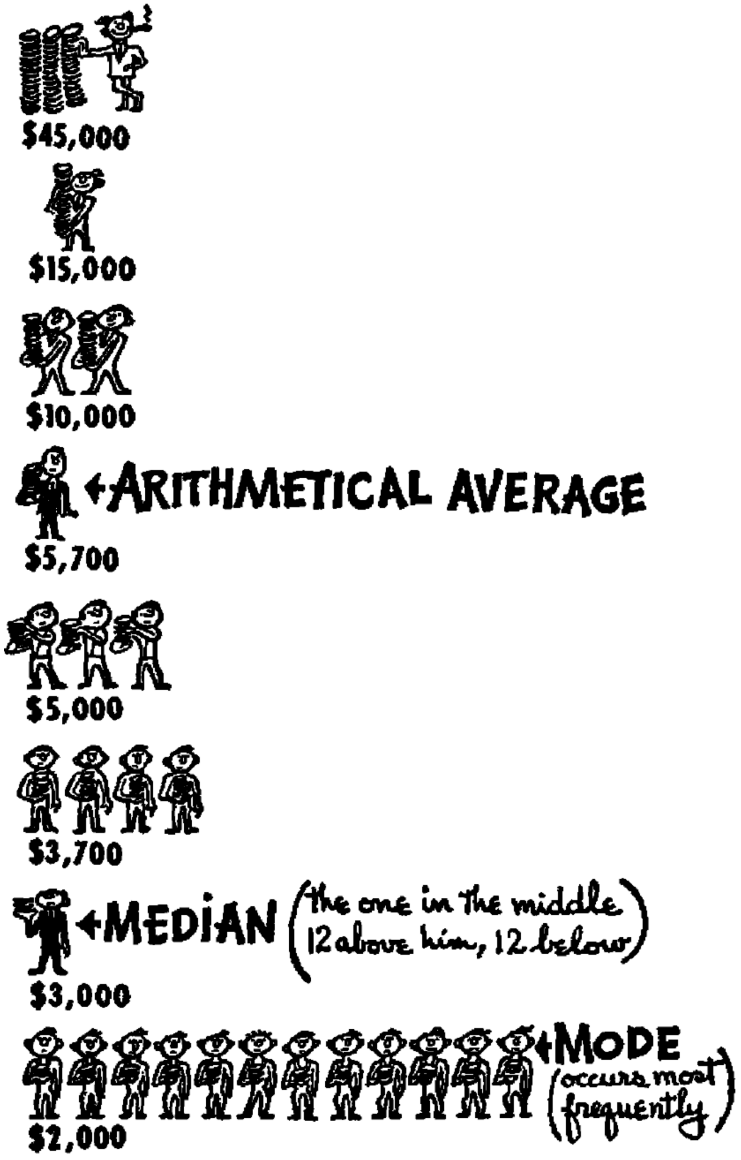
Different kinds of averages

Sample mean — arithmetic mean of the sample.
Sample median — central element of the sorted sample.
Sample mode — the most common value in the sample.

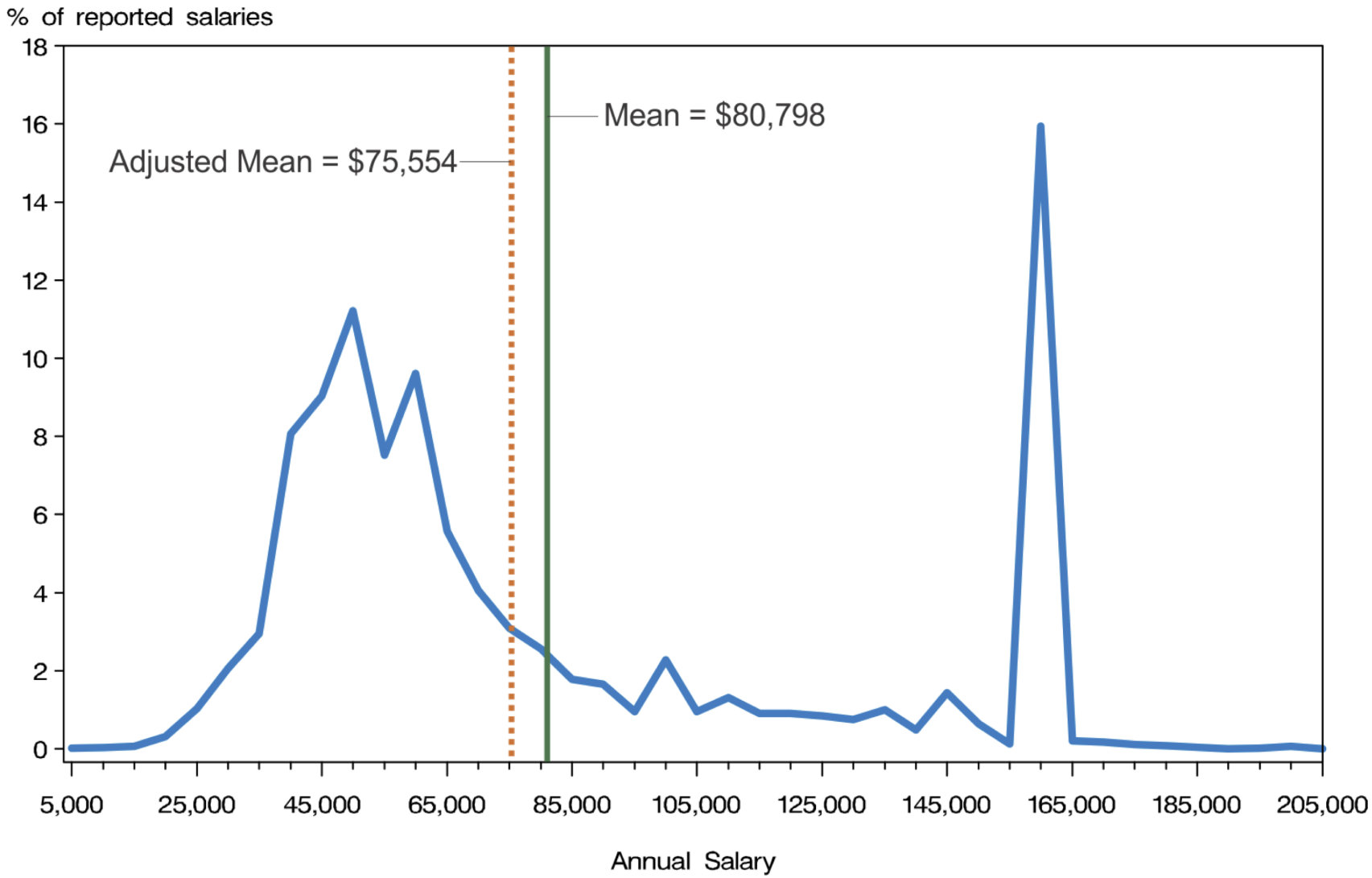


Different kinds of averages

(Huff, 1954):



Limitations of statistics



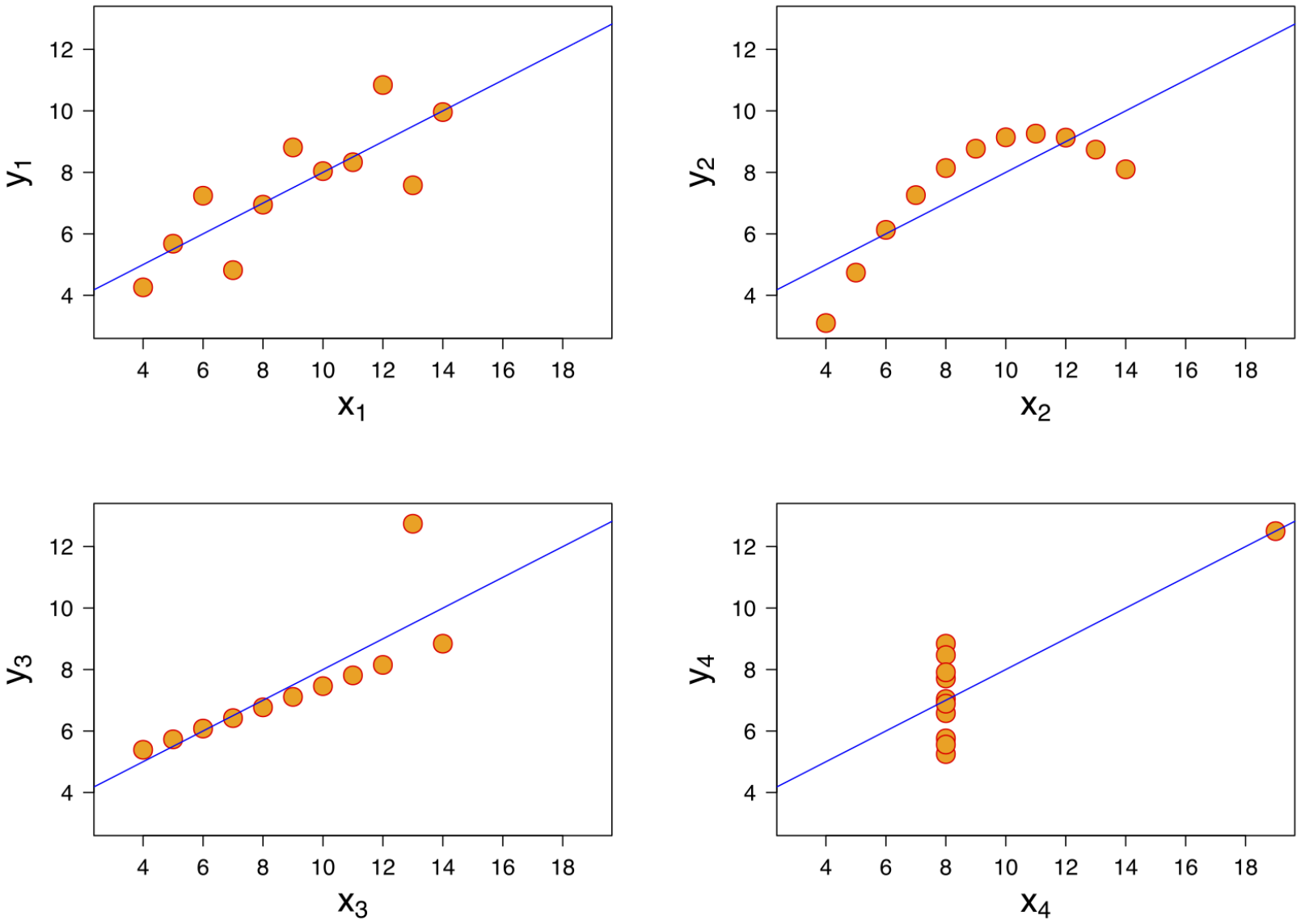
Distribution of reported full-time salaries of new law graduates, NALP, USA, 2012.

Limitations of statistics

Anscombe's quartet (Anscombe, 1973):

Number	1	2	3	4
$\bar{x}$	9	9	9	9
S_x	11	11	11	11
$\bar{y}$	7.5	7.5	7.5	7.5
S_y	4.127	4.127	4.128	4.128
r_{xy}	0.816	0.816	0.816	0.816

Limitations of statistics



Point estimates

Let the population distribution be parametric:

$$F(x) = F(x, \theta).$$

Statistic $\hat{\theta}_n = \hat{\theta}\left(X^n\right)$ is a point estimate of the parameter θ.
Which estimate is better?

Consistency: $\operatorname*{plim}_{n\to\infty} \hat{\theta}_n = \theta$.
Unbiasedness: $\mathbb{E}\hat{\theta}_n = \theta$.
Asymptotic unbiasedness: $\lim_{n\to\infty} \mathbb{E}\hat{\theta}_n = \theta$.
Optimality: $\mathbb{D}\hat{\theta}_n = \min_{\hat{\theta}\,:\,\mathbb{E}\hat{\theta}=\theta} \mathbb{D}\hat{\theta}$.
Robustness: resistance of $\hat{\theta}_n$ to:

- deviations of the true population distribution of X from the model family $F(x, \theta)$
- outliers in the sample

Maximum likelihood estimation

A popular method of point estimation:

$$
\begin{aligned}
X &\sim f(x, \theta), \\
X^n &= (X_1, \ldots, X_n), \\
L(X^n, \theta) &= \prod_{i=1}^{n} f(X_i, \theta), \\
\hat{\theta}_{MLE} &\equiv \operatorname*{argmax}_{\theta} L(X^n, \theta).
\end{aligned}
$$

It's convenient to take a logarithm:

$$
\begin{aligned}
\log L(X^n, \theta) &= \sum_{i=1}^{n} f(X_i, \theta), \\
\hat{\theta}_{MLE} &\equiv \operatorname*{argmax}_{\theta} \log L(X^n, \theta).
\end{aligned}
$$

Derivatives of likelihood function

Score function:

$$S(\theta) \equiv \frac{\partial}{\partial \theta} \log L(\theta)$$

MLE is a solution to score equation:

$$S(\theta) = 0$$

Fisher's information:

$$I(\theta) \equiv -\frac{\partial^2}{\partial \theta^2} \log L(\theta)$$

Variance of MLE:

$$\mathbb{D}\hat{\theta}_{MLE} \approx I^{-1}\left(\hat{\theta}_{MLE}\right)$$

Properties of MLEs

- consistency:

$$\operatorname*{plim}_{n\to\infty} \hat{\theta}_{MLE} = \theta$$

- asymptotic normality: with $n \to \infty$

$$\hat{\theta}_{MLE} \sim N\left(\theta, I^{-1}(\theta)\right)$$

- efficiency: MLE has the smallest variance of all the unbiased estimates
- invariance: $g\left(\hat{\theta}_{MLE}\right)$ — MLE for $g(\theta)$

Interval estimates

Confidence interval:

$$\mathbf{P}(\theta \in [C_L, C_U]) \geqslant 1 - \alpha,$$

$1 - \alpha$ — confidence level,
C_L, C_U — lower and upper confidence limits.

Wrong interpretation: the unknown parameter is inside the obtained confidence interval with probability $1 - \alpha$.

Correct interpretation: if the procedure of building a confidence interval is repeated infinitely on random samples from the same population, $100(1 - \alpha)$% of the intervals would contain the true value of θ.

For normal distribution

$X \sim N\left(\mu, \sigma^2\right), \quad X^n = (X_1, \ldots, X_n),$

$\bar{X}_n$ — an estimate of $\mathbb{E}X = \mu$,

CLT: $\bar{X}_n \sim N\left(\mu, \frac{\sigma^2}{n}\right) \Rightarrow$

$$\mathbf{P}\left(\mu - z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \leqslant \bar{X}_n \leqslant \mu + z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha \Rightarrow$$

confidence interval for μ:

$$\mathbf{P}\left(\bar{X}_n - z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \leqslant \mu \leqslant \bar{X}_n + z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha,$$

$z_{1-\frac{\alpha}{2}}$ — a quantile of the standard normal distribution.

For non-normal distributions

CLT: if X^n is a sample from $F(x)$, $F(x)$ is not too skewed, and $n > 30$, then

$$\bar{X}_n \sim\approx N\left(\mathbb{E}X, \frac{\mathbb{D}X}{n}\right) \Rightarrow$$

confidence interval for $\mathbb{E}X$:

$$\mathbf{P}\left(\bar{X}_n - z_{1-\frac{\alpha}{2}}\sqrt{\frac{\mathbb{D}X}{n}} \leqslant \mathbb{E}X \leqslant \bar{X}_n + z_{1-\frac{\alpha}{2}}\sqrt{\frac{\mathbb{D}X}{n}}\right) \approx 1-\alpha.$$

If the variance is unknown:

$$\mathbf{P}\left(\bar{X}_n - t_{n-1,1-\frac{\alpha}{2}}\frac{S_n}{\sqrt{n}} \leqslant \mathbb{E}X \leqslant \bar{X}_n + t_{n-1,1-\frac{\alpha}{2}}\frac{S_n}{\sqrt{n}}\right) \approx 1-\alpha,$$

$t_{n-1,1-\frac{\alpha}{2}}$ — quantile of the Student's t-distribution with $n-1$ degrees of freedom.

Quantiles

Nonparametric confidence interval for the median of the continuous distribution:

$$X^n = (X_1, \ldots, X_n), \quad X \sim F(x) \Rightarrow$$

$$\mathbf{P}\left(\operatorname{med} X \in \left[X_{(r)}, X_{(n-r+1)}\right]\right) = \frac{1}{2^n} \sum_{i=r}^{n-r+1} C_n^i.$$

For $n > 10$ normal approximation might be used:

$$\mathbf{P}\left(\operatorname{med} X \in \left[X_{\left(\left\lfloor \frac{n-\sqrt{n}z_{1-\frac{\alpha}{2}}}{2} \right\rfloor\right)}, X_{\left(\left\lceil \frac{n+\sqrt{n}z_{1-\frac{\alpha}{2}}}{2} \right\rceil\right)}\right]\right) \approx 1 - \alpha.$$

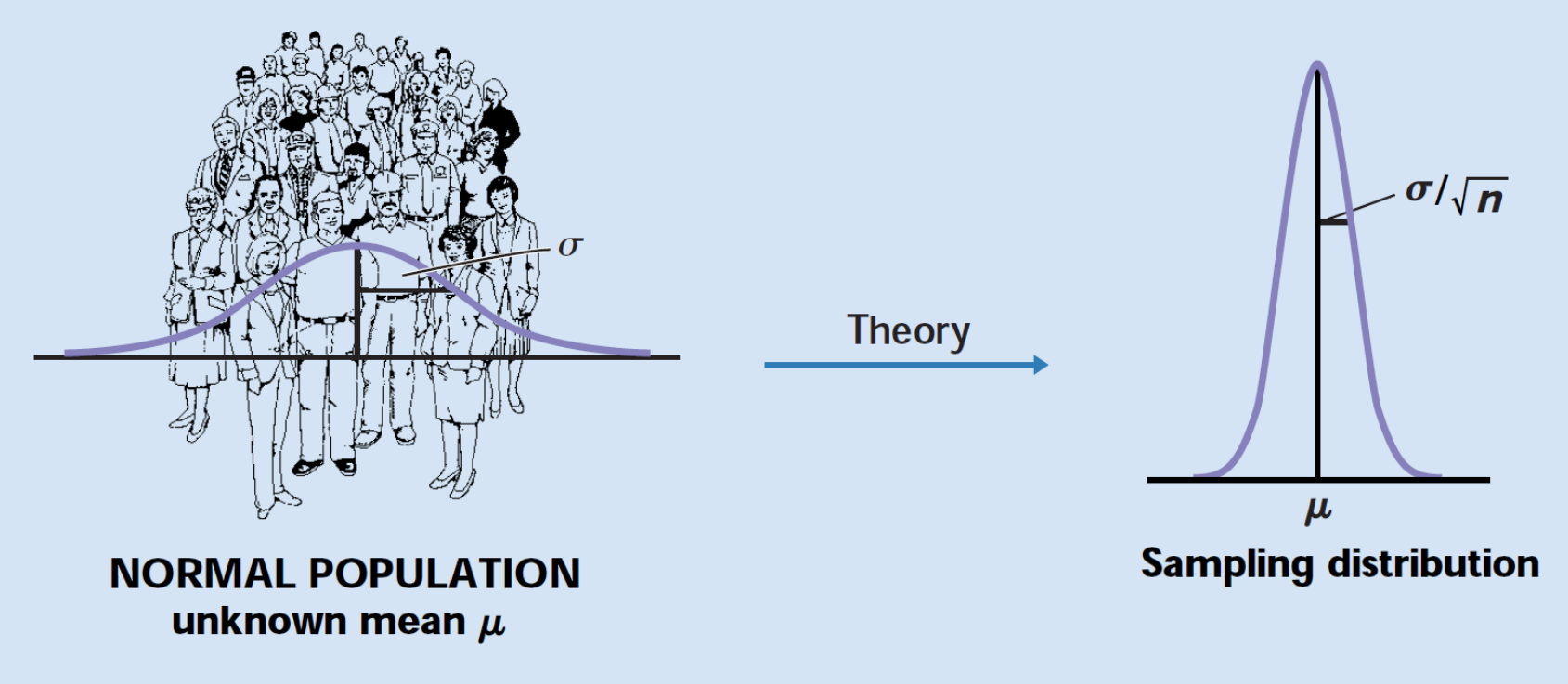
Same for any $\alpha \in (0, 1)$:

$$\mathbf{P}\left(X_\alpha \in \left[X_{(l)}, X_{(u)}\right]\right) = \sum_{i=l}^{u} C_n^i \alpha^i (1-\alpha)^{n-i}.$$

Building confidence intervals

How can we estimate $F_{\hat{\theta}_n}(x)$ — sampling distribution of the statistic $\hat{\theta}_n$? (Hesterberg, 2005):

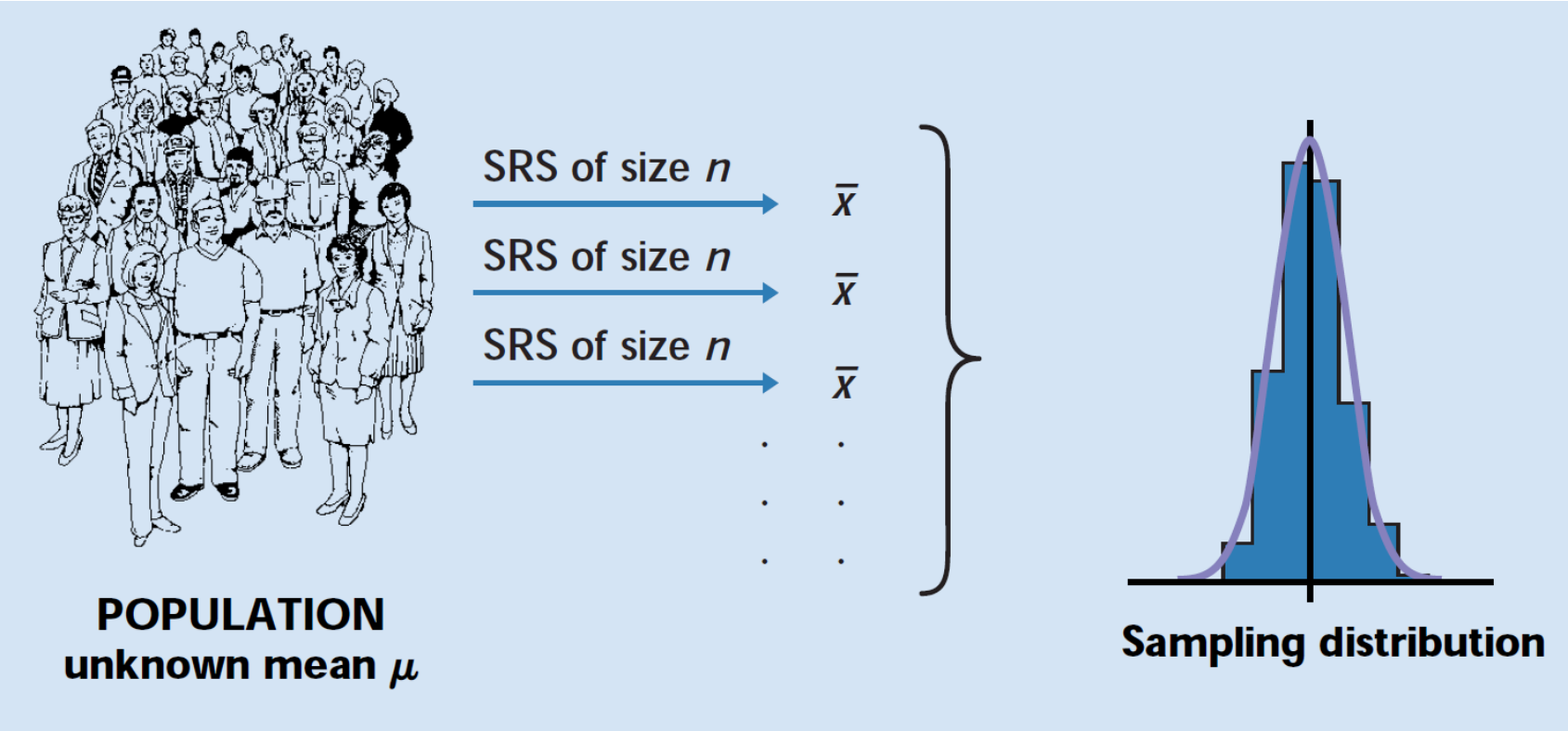
- parametric method:



Assume X is distributed according to $F_X(x)$, for which the distribution of $\hat{\theta}_n$ could be analytically inferred.

Building confidence intervals

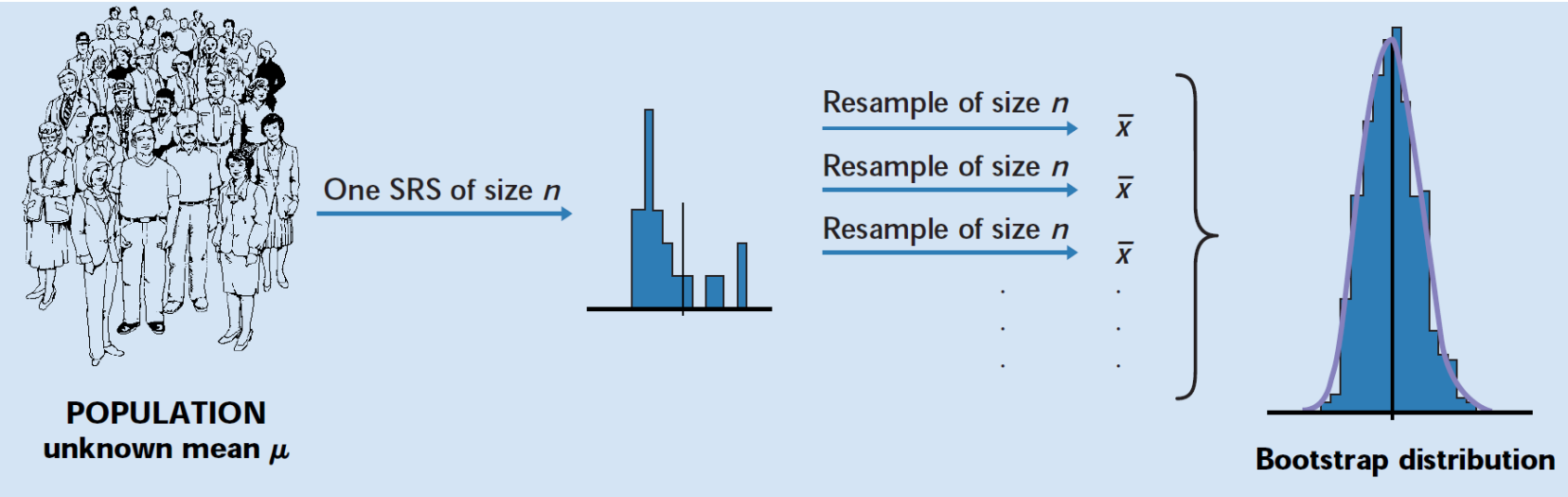
- naive method:



Take N samples of size n from the population, estimate the distribution of $\hat{\theta}_n$ by its EDF.

Building confidence intervals

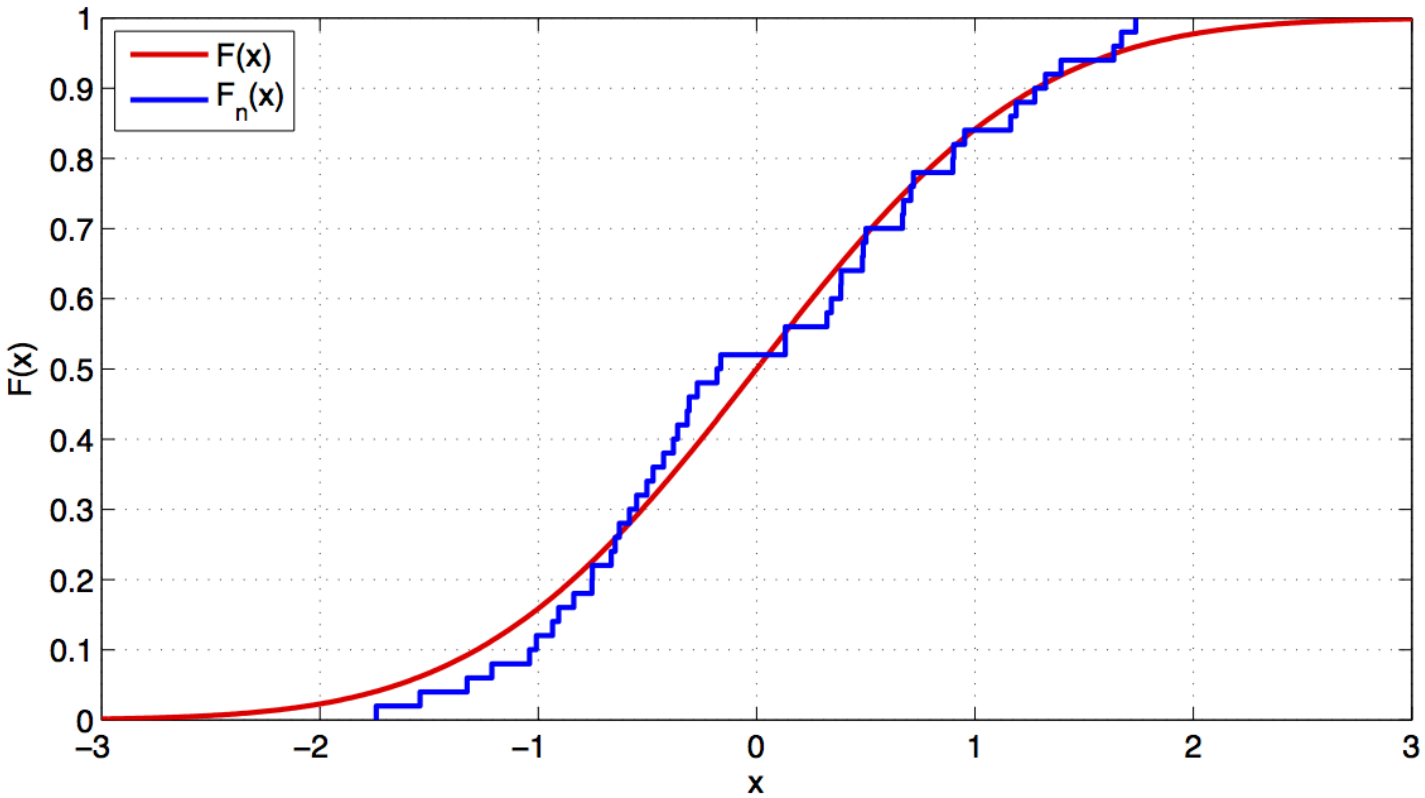
- bootstrap:



Take N resamples (with repetition) of size n from the original sample, estimate the distribution of $\hat{\theta}_n$ by its EDF on resamples.

Bootstrap

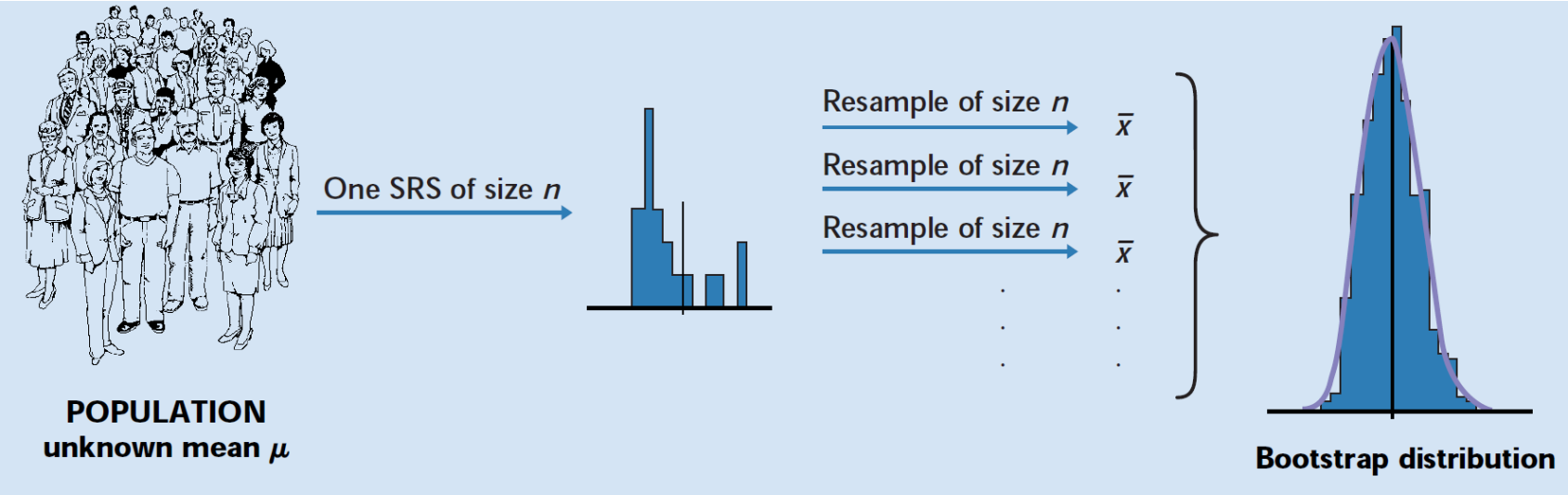
Sampling from population (as in naive method) means sampling from the unknown distribution $F_X(x)$.
Best estimate of $F_X(x)$ we have is $F_{X^n}(x)$:



Sampling from $F_{X^n}(x)$ is the same as taking resamples from X^n.

Bootstrap distribution

$X^{1*}, \ldots, X^{N*}$ — resamples from X^n of size n,
$\hat{\theta}_n^{1*}, \ldots, \hat{\theta}_n^{N*}$ — values of the statistic on them,
$F_{\hat{\theta}_n}^{boot}(x)$ — bootstrap distribution of $\hat{\theta}_n$ — EDF of values of statistic on resamples.



With $F_{\hat{\theta}_n}^{boot}(x)$ we can build confidence intervals for θ!

Bootstrap confidence intervals

- Calculate S_n^{boot} — sample standard deviation of $\hat{\theta}_n$ on resamples;

$$\mathbf{P}\left(\hat{\theta}_n - t_{n-1,1-\frac{\alpha}{2}} S_n^{boot} \leqslant \theta \leqslant \hat{\theta}_n + t_{n-1,1-\frac{\alpha}{2}} S_n^{boot}\right) \approx 1 - \alpha.$$

  This is studentized bootstrap.

- Take sample quantiles of the bootstrap distribution:

$$\mathbf{P}\left(\left(F_{\hat{\theta}_n}^{boot}\right)^{-1}\left(\frac{\alpha}{2}\right) \leqslant \theta \leqslant \left(F_{\hat{\theta}_n}^{boot}\right)^{-1}\left(1 - \frac{\alpha}{2}\right)\right) \approx 1 - \alpha.$$

  This is basic bootstrap.

Bootstrap confidence intervals

- Tweak basic bootstrap a little:

$$\mathbf{P}\left(\left(F^{boot}_{\hat{\theta}_n}\right)^{-1}(\alpha_1) \leqslant \theta \leqslant \left(F^{boot}_{\hat{\theta}_n}\right)^{-1}(\alpha_2)\right) \approx 1-\alpha,$$

$$\alpha_1 = \Phi\left(\hat{z}_0 + \frac{\hat{z}_0 + z_{\frac{\alpha}{2}}}{1-\hat{a}\left(\hat{z}_0 + z_{\frac{\alpha}{2}}\right)}\right),$$

$$\alpha_2 = \Phi\left(\hat{z}_0 + \frac{\hat{z}_0 + z_{1-\frac{\alpha}{2}}}{1-\hat{a}\left(\hat{z}_0 + z_{1-\frac{\alpha}{2}}\right)}\right),$$

$$\hat{z}_0 = \Phi^{-1}\left(\frac{1}{N}\sum_{i=1}^{N}\left[\hat{\theta}^{i*}_n < \hat{\theta}_n\right]\right),$$

$\hat{a}$ too big to fit on this slide.

This is bias-corrected accelerated bootstrap.

Bootstrap estimates

- are asymptotically consistent
- are easy to construct even for complex statistics
- have poor performance for statistics that depend on a small number of observations (like median)

Hypothesis testing

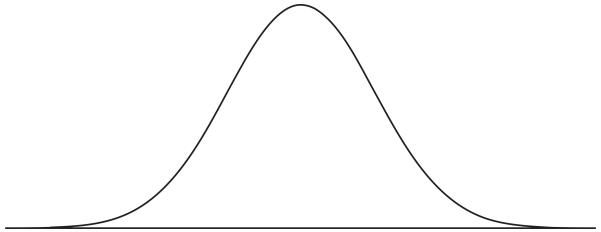
sample: $X^n = (X_1, \ldots, X_n)\,,\ X \sim \mathbf{P} \in \Omega$

null hypothesis: $H_0\colon\ \mathbf{P} \in \omega,\ \ \omega \in \Omega$

alternative hypothesis: $H_1\colon\ \mathbf{P} \notin \omega$

statistic: $T(X^n)\,,\ \ T(X^n) \sim F(x)$ when $\mathbf{P} \in \omega$

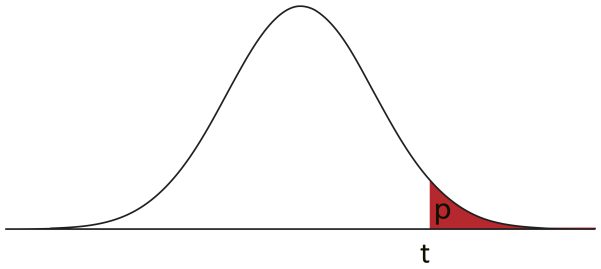
$T(X^n) \not\sim F(x)$ when $\mathbf{P} \notin \omega$



observed sample: $x^n = (x_1, \ldots, x_n)$

observed statistic: $t = T(x^n)$

p-value: $p(x^n)$ — probability of getting $T(X^n) = t$ or more extreme under H_0



$$p(x^n) = \mathbf{P}(T \geqslant t \,|\, H_0)$$

Null is rejected when $p(x^n) \leqslant \alpha,\ \ \alpha$ — significance level

Hypothesis testing

P-value

$$p = \mathbf{P}(T \geqslant t \,|H_0\,) \neq \mathbf{P}(H_0)$$

Example: it was claimed that Paul the Octopus was able to predict the results of football matches by choosing the feeder with the flag of the country that was about to win. Out of 14 predictions, 12 were correct. Testing the null hypothesis that he "guesses"at random, by an appropriate statistical test we get $p \approx 0.0129$.

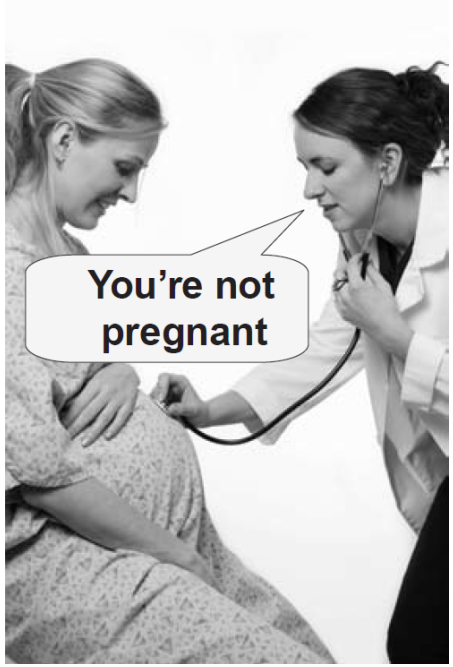
0.0129 is *not* the probability that the octopus is guessing at random! It equals to one.

Type I and II errors

	H_0 true	H_0 false
H_0 accepted	H_0 correctly accepted	Type II error (False negative)
H_0 rejected	Type I error (False positive)	H_0 correctly rejected

Type I error
(false positive)

Type II error
(false negative)



Type I and II errors

Hypothesis testing is not symmetrical regarding the pair (H_0, H_1): the probability of type I error is bounded from above by α, and of type II error — minimized by using more suitable statistical test.

Correct test: $\mathbf{P}(p\,(T) \leqslant \alpha\,|H_0\,) \leqslant \alpha \forall\, \mathbf{P} \in \Omega$.
Power: $\mathrm{pow} = \mathbf{P}(p\,(T) \leqslant \alpha\,|H_1\,)\,.$
Consistent test: $\mathrm{pow} \to 1$ with $n \to \infty$ for every untrue H_0.
T_1 — **uniformly most powerful** test, if $\forall T_2$

$$\mathbf{P}(p\,(T_1) \leqslant \alpha\,|H_1\,) \geqslant \mathbf{P}(p\,(T_2) \leqslant \alpha\,|H_1\,) \quad \forall H_1 \neq H_0,$$
$$\mathbf{P}(p\,(T_1) \leqslant \alpha\,|H_0\,) = \mathbf{P}(p\,(T_2) \leqslant \alpha\,|H_0\,)\,,$$

and for at least one H_1 the inequality is strict.

Interpretation

Low p — evidence against null hypothesis in favour of the alternative.
If p is not low enough, there is no evidence against null hypothesis in favour of the alternative.

Absence of evidence $\not\Rightarrow$ evidence of absence.
Using statistical tests, it is impossible to confirm null hypothesis!

Statistical and practical significance

The probability of rejecting wrong null hypothesis depends not only on how far it is from the truth, but on the sample size as well.
The bigger the sample size n, the more subtle departures from H_0 our statistic detects.

When testing hypothesis, one should always estimate an **effect size** — the magnitude of departure from null, and evaluate its practical significance.

Statistical and practical significance

- (Lee et al, 2010): after 3 years women who exercised more than 1 hour per day, on average gained significantly less weight than women who exercised less than 20 minutes per day ($p < 0.001$). The difference was 150 g. Practical significance of this is questionable. Details: `http://youtu.be/oqDZO-mfN4Q`.
- (Ellis, 2010, chapter 2): in 2002 a clinical trial of Premarin, a hormonal treatment of the menopausal syndromes, was terminated. It was discovered that its usage leads to significant increase of the risks of breast cancer (0.08% up), stroke (0.08%), and heart attack (0.07%). The effects look small, but not when multiplied by the population size — it's thousands additional deaths.
- (Kirk, 1996): if during the trial of the hypothetical drug that is supposed to slow down the progression of the Alzheimer's disease the effect size — difference in average IQ loss between treatment and control groups — is found to be 13 points, maybe the trial should be continued even if the difference is not statistically significant.

Other details

- Chosen statistic might not reflect all the relevant information from the sample. Example:

$$H_0 \colon X \sim N\left(\mu, \sigma^2\right), \quad H_1 \colon H_0 \text{ is false;}$$

$$T\left(X^n\right) = g_1.$$

  For all symmetric non-normal distribution we would fail to reject the null.
- Point hypotheses $H_0 \colon \theta = \theta_0$ could be tested with confidence intervals for θ:
  - if θ_0 is not inside the $100\,(1-\alpha)$% confidence interval for θ, then H_0 is rejected at level α;
  - p-value — maximal α for which θ_0 falls inside the corresponding confidence interval.

Shaken, not stirred

James Bond claims to prefer his martini shaken, not stirred. Let's run a blind test: offer him n pairs to choose from to see which would he prefer more often.

Sample: binary vector of length n, 1 — James Bond chooses shaken, 0 — stirred martini.

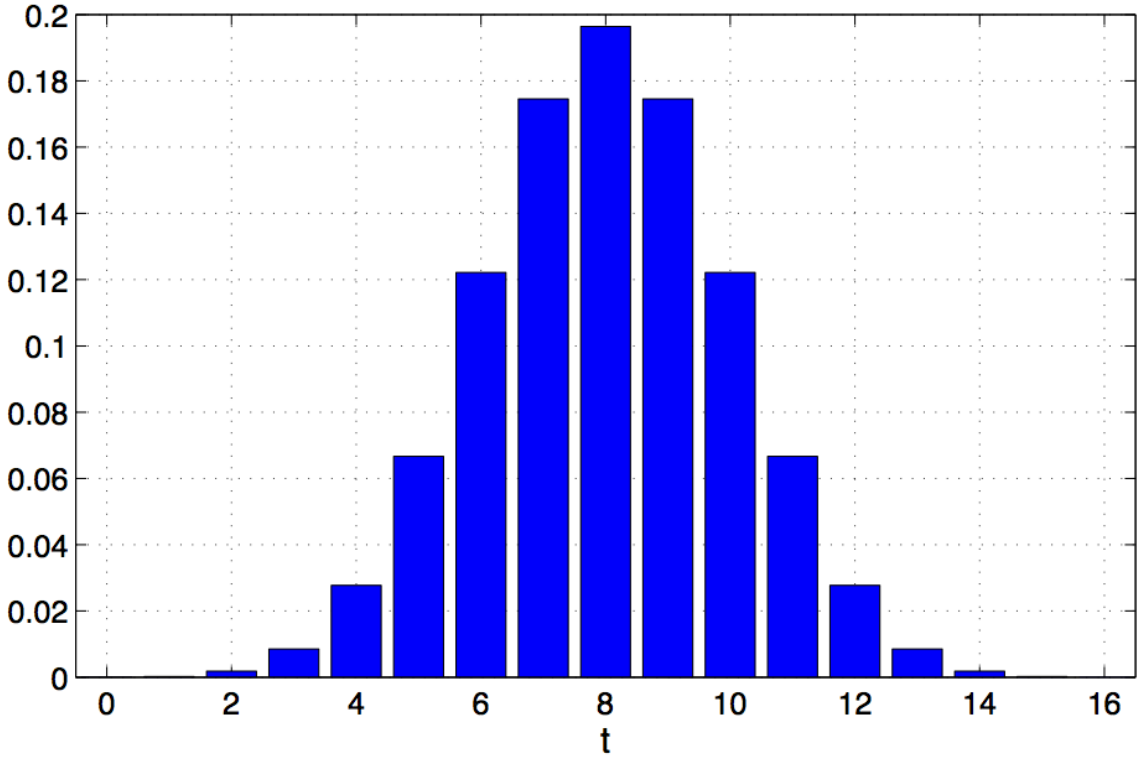
H_0: the subject is not able to differentiate between two types of martini and selects at random.

Statistic T — number of ones in the sample.

Null distribution

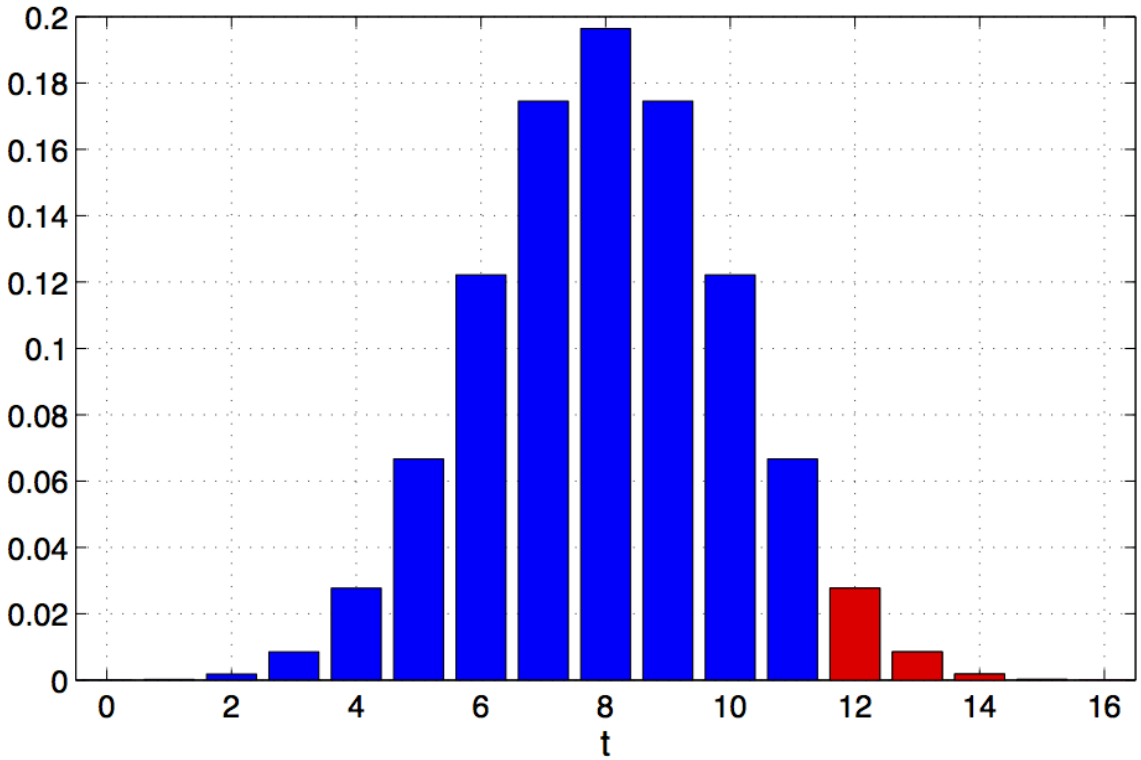
If the null hypothesis is true, when all possible samples of size n of zeroes and ones have the same probability.

Let $n = 16$, then there are $2^{16} = 65536$ equiprobable samples. Statistic T takes values between 0 and 16:



One-sided alternative

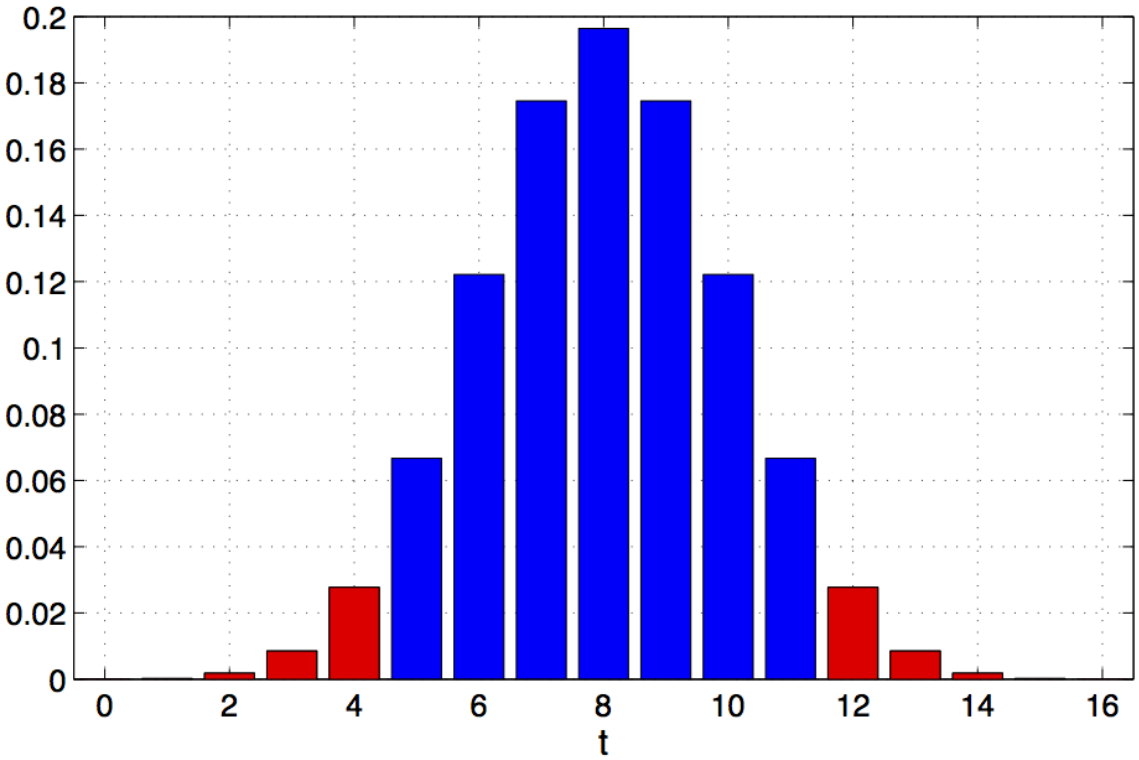
H_1: the subject prefers shaken martini.
If that is true, bigger values of T are more likely (i.e., large T is an evidence against H_0 for H_1).
Under H_0 the probability that James Bond would select shaken martini 12 or more of 16 trials is equal to $\frac{2517}{65536} \approx 0.0384$.



0.0384 — p-value for $t = 12$.

Two-sided alternative

H_1: the subject prefers one of the types of martini.
If that is true, both large and small values of T are evidence against H_0 for H_1).
Under H_0 he probability that James Bond would select shaken martini $\geqslant 12$ or $\leqslant 4$ of 16 trials equals $\frac{5034}{65536} \approx 0.0768$.



0.0768 — p-value for $t = 12$.

P-value

The lower the p-value, the stronger the evidence for the alternative against null.

0.0384 — a probability of observing $t \geqslant 12$ under true null, i. e. the subject select drinks at random.

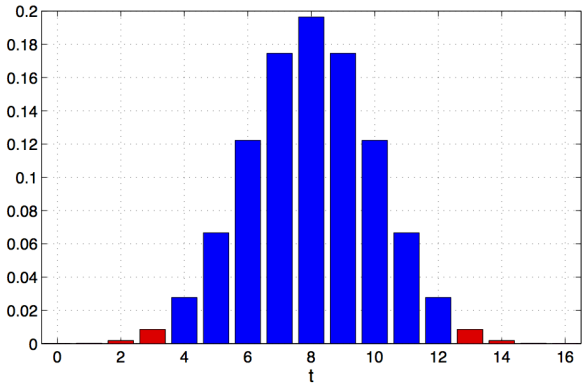
Again: it is not the probability that the null hypothesis is true!

Example: imagine that James Bond selects shaken martini in 51% of cases (unobserved probability).

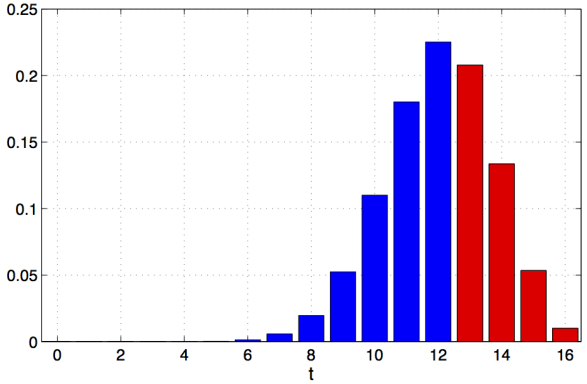
If after 100 trials shaken martini was selected 49 times, for the one-sided alternative $p \approx 0.6178$. The null is not rejected, but it's not true — the subject has a certain preference!

Power

When testing the null against two-sided alternative, we reject when $t \geqslant 13$ or $t \leqslant 3$. The borderline values of the statistic gives $p = 0.0213 \leqslant \alpha = 0.05$.



Assume James Bond selects shaken martini in 75% of cases (unobserved probability).



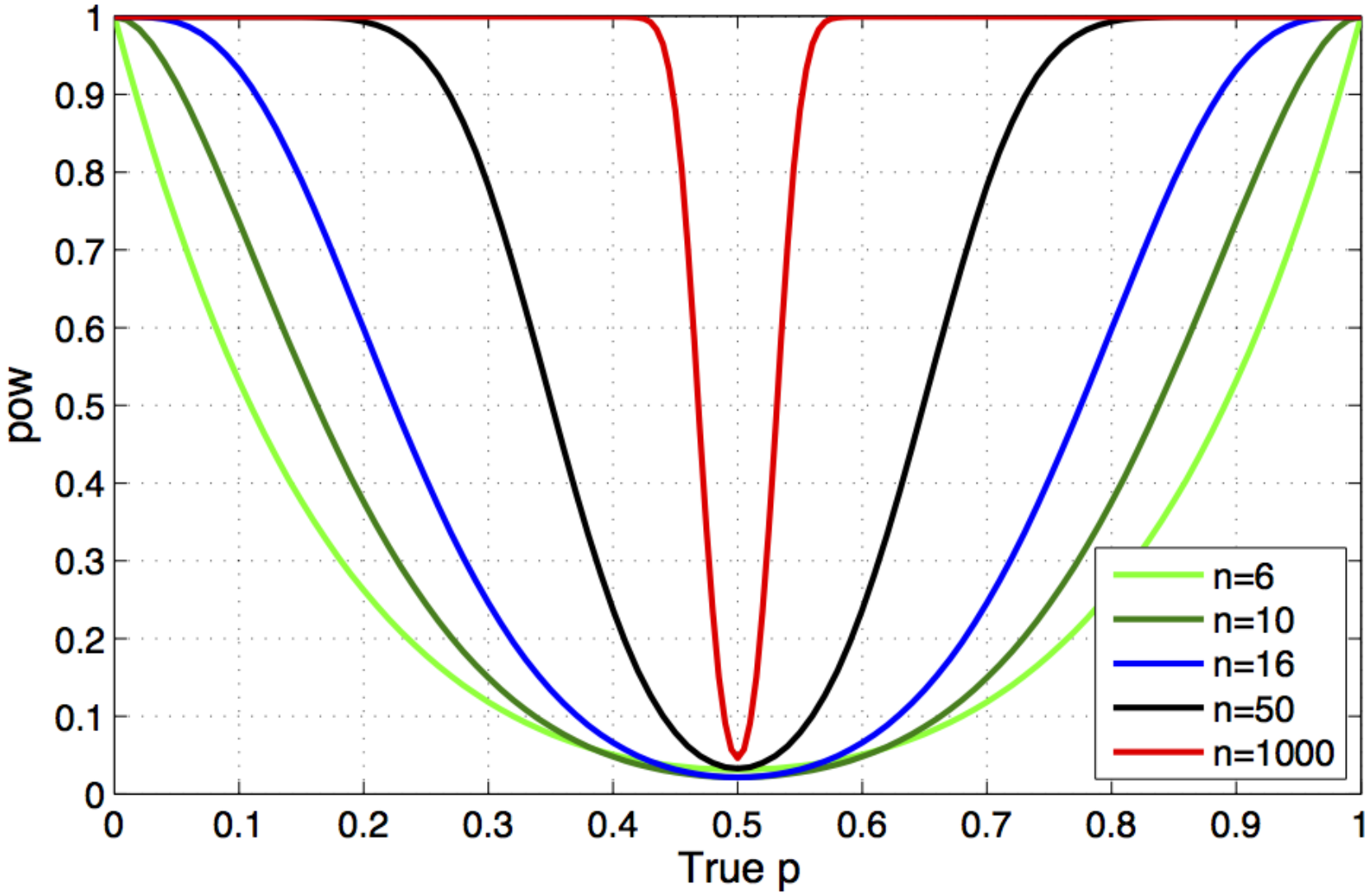
$\mathrm{pow} \approx 0.6202$, i. e., if we repeat the whole experiment (infinitely) many times, the null would only be rejected in 62% of experiments.

Power

Power of the test depends on:

- sample size
- effect size
- sensitivity of the statistic
- type of the alternative

Power



Sample size

What sample size do we need?

Real-world limitations: 1 martini contains about 25 ml of alcohol. A lethal dose for 80 kg person could be 320-960 ml of alcohol depending on the tolerance (13-38 martinis).

Providing sufficient power: sample size is selected so that for the smallest practically interesting effect size (for example, shaken martini preference probability 0.75) the power is above some level (say, 0.80).

References

A handbook on hypothesis testing:

- Kanji G.K. *100 statistical tests*, 2006.

Introductory textbooks:

- Good P.I., Hardin J.W. *Common Errors in Statistics (and How to Avoid Them)*, 2003.
- Reinhart A. *Statistics Done Wrong. The woefully complete guide*, `http://www.statisticsdonewrong.com/`

R:

- a video with Trevor Hastie: `http://youtu.be/jwBgGS_4RQA`
- short reference card: `http://cran.r-project.org/doc/contrib/Short-refcard.pdf`
- learn R in R: `http://swirlstats.com`

References

Likelihood:

- Chihara L., Hesterberg T. *Mathematical Statistics with Resampling and R*, 2011, chapter 6.
- Pawitan, Y. *In all likelihood: statistical modelling and inference using likelihood*, 2001.

Bootstrap:

- Hesterberg T., Monaghan S., Moore D.S., Clipson A., Epstein R. *Bootstrap methods and permutation tests*. In Introduction to the Practice of Statistics, 2005. `http://statweb.stanford.edu/~tibs/stat315a/Supplements/bootstrap.pdf`
- Efron B., Tibshirani R. *An Introduction to the Bootstrap*, 1993.

Hypothesis testing:

- Good P.I., Hardin J.W. *Common Errors in Statistics (and How to Avoid Them)*, 2003, chapter 2.

References

Anscombe F.J. (1973). *Graphs in Statistical Analysis*. American Statistician, 27(1): 17–21.

Ellis P.D. *The Essential Guide to Effect Sizes: Statistical Power, Meta-Analysis, and the Interpretation of Research Results*, 2010.

Huff D. *How To Lie With Statistics*, 1954.

Kahneman D., Frederick S. (2002). *Representativeness Revisited: Attribute Substitution in Intuitive Judgment*. In T. Gilovich, D. Griffin, D. Kahneman (Eds.), Heuristics and biases: The psychology of intuitive judgement (pp. 49–81).

Kirk R.E. (1996). *Practical Significance: A Concept Whose Time Has Come*. Educational and Psychological Measurement, 56(5), 746–759.

Lee I.-M., Djoussè L., Sesso H.D., Wang L., Buring J.E. (2010). *Physical Activity and Weight Gain Prevention*. JAMA: the Journal of the American Medical Association, 303(12), 1173–1179.

Marriott, F. H. C. *The Interpretation of Multiple Observations*, 1974.