Probability and Statistics Overview

Lecture-3

Slides have been prepared from multiple sources on the Internet. No authorship is claimed.

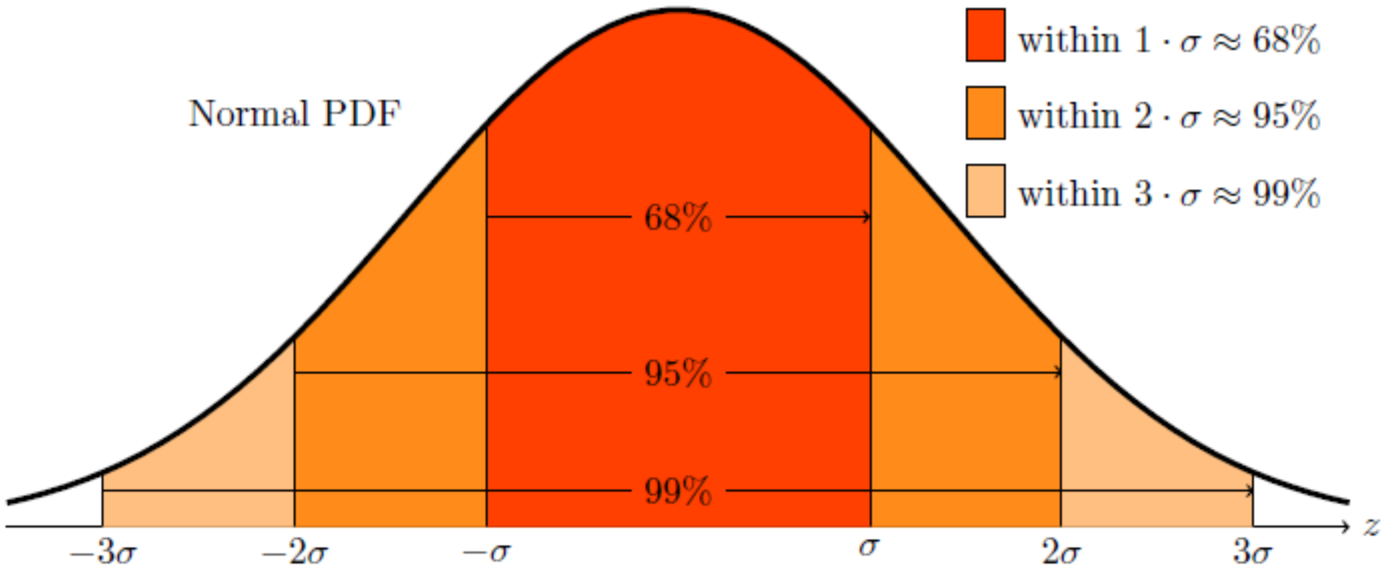
Standardization

Random variable X with mean μ and standard deviation σ.

Standardization: $Y = \dfrac{X - \mu}{\sigma}$.

- Y has mean 0 and standard deviation 1.
- Standardizing any normal random variable produces the standard normal.
- If $X \approx$ normal then standardized $X \approx$ stand. normal.
- We use reserve Z to mean a standard normal random variable.

Concept Question: Standard Normal



1. $P(-1 < Z < 1)$ is

(a) 0.025 (b) 0.16 (c) 0.68 (d) 0.84 (e) 0.95

2. $P(Z > 2)$

(a) 0.025 (b) 0.16 (c) 0.68 (d) 0.84 (e) 0.95

answer: 1c, 2a

Central Limit Theorem

Setting: $X_1, X_2, \ldots$ i.i.d. with mean μ and standard dev. σ.

For each n:

$$\overline{X}_n = \frac{1}{n}(X_1 + X_2 + \ldots + X_n) \qquad \text{average}$$

$$S_n = X_1 + X_2 + \ldots + X_n \qquad \text{sum.}$$

Conclusion: For large n:

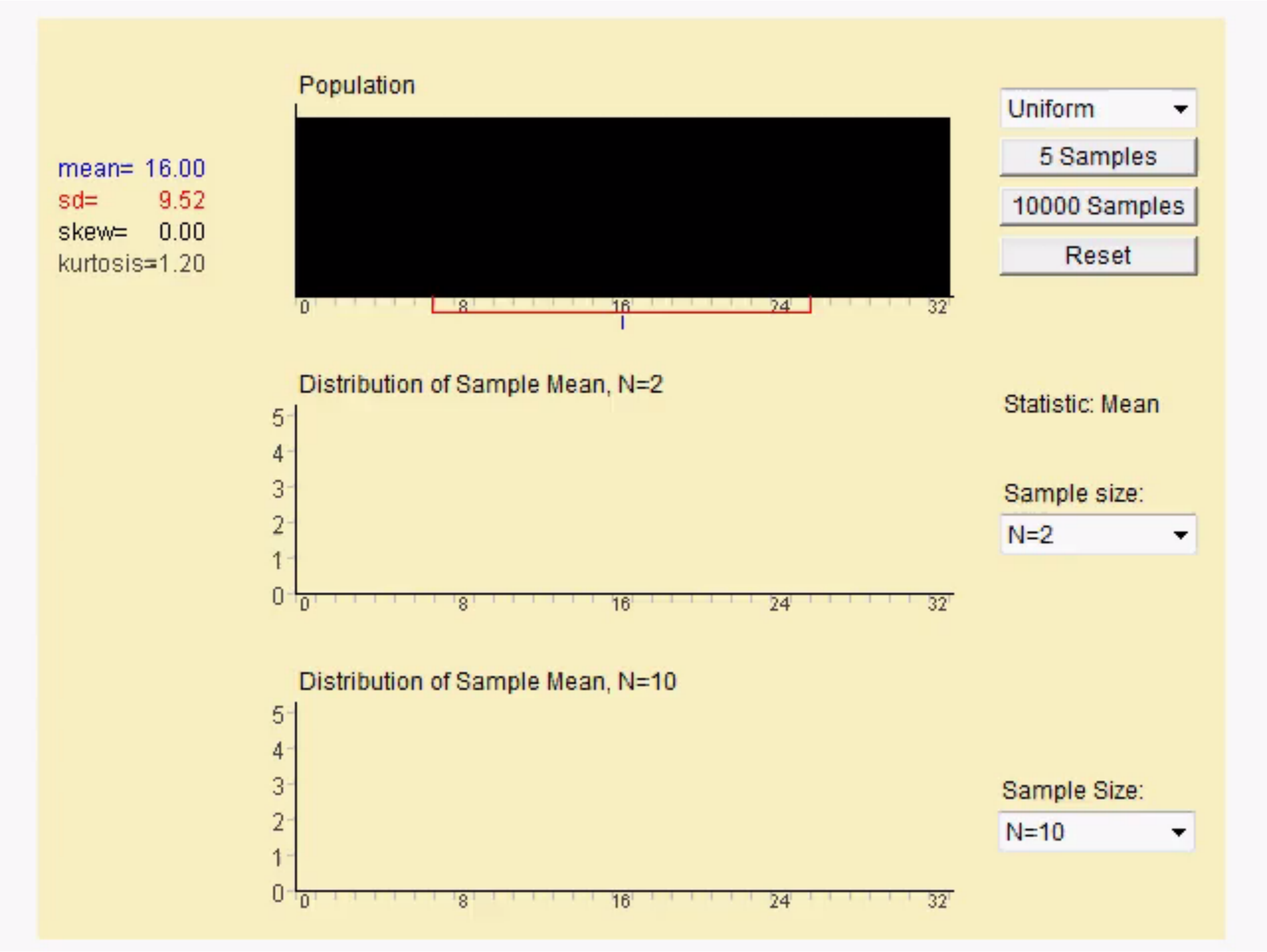
$$\overline{X}_n \approx \mathrm{N}\left(\mu, \frac{\sigma^2}{n}\right)$$

$$S_n \approx \mathrm{N}\left(n\mu, n\sigma^2\right)$$

Standardized S_n or $\overline{X}_n \approx \mathrm{N}(0,1)$

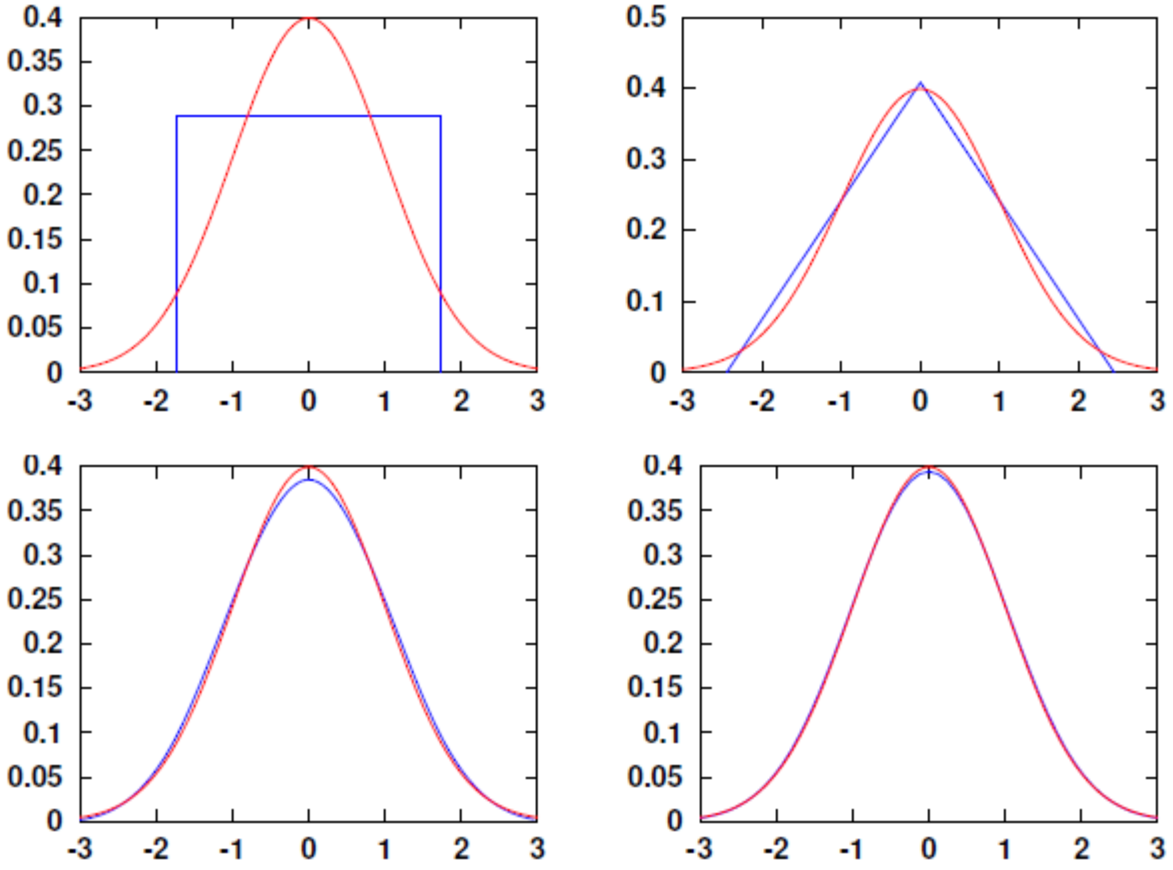
That is, $$\frac{S_n - n\mu}{\sqrt{n}\sigma} = \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \approx \mathrm{N}(0,1).$$

http://onlinestatbook.com/2/sampling_distributions/clt_demo.html


mean= 16.00
sd= 9.52
skew= 0.00
kurtosis=1.20
Population
0 8 16 24 32
Uniform
5 Samples
10000 Samples
Reset
Distribution of Sample Mean, N=2
0 1 2 3 4 5
0 8 16 24 32
Statistic: Mean
Sample size:
N=2
Distribution of Sample Mean, N=10
0 1 2 3 4 5
0 8 16 24 32
Sample Size:
N=10

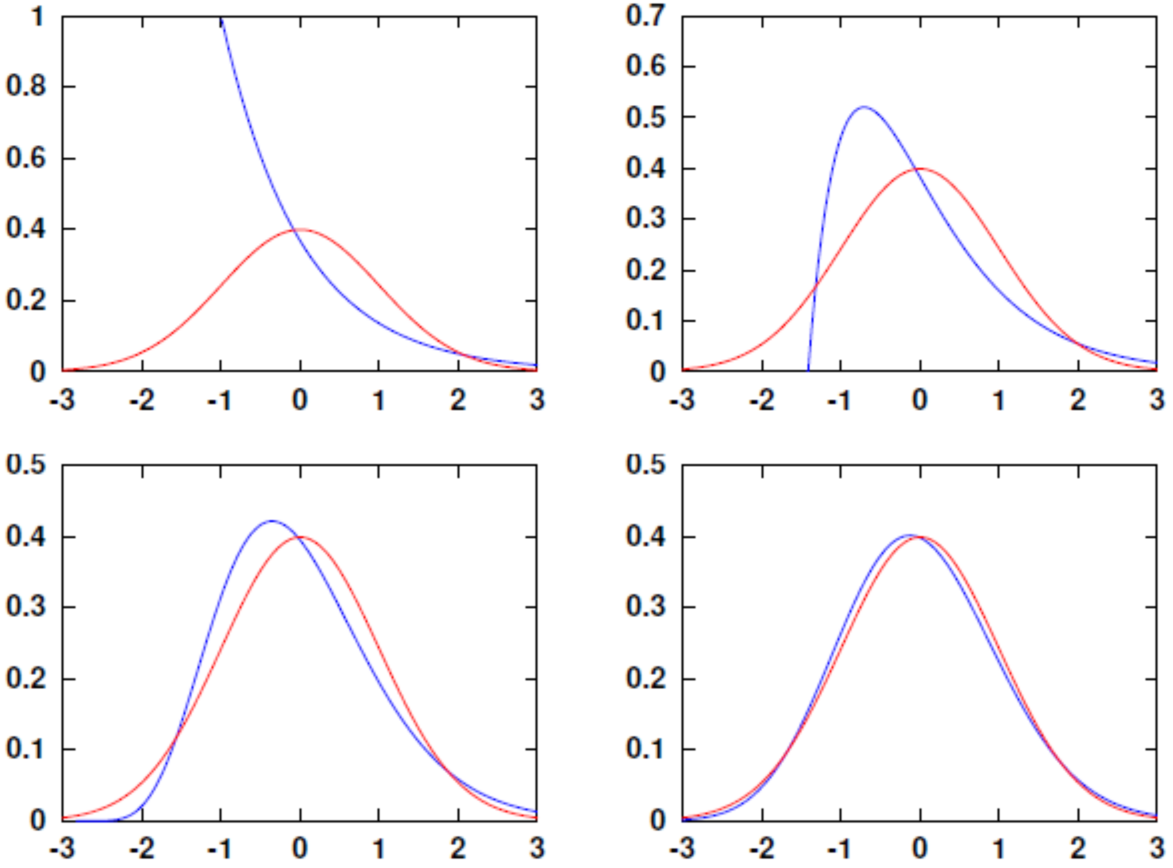
CLT: pictures

Standardized average of n i.i.d. uniform random variables with $n = 1, 2, 4, 12$.



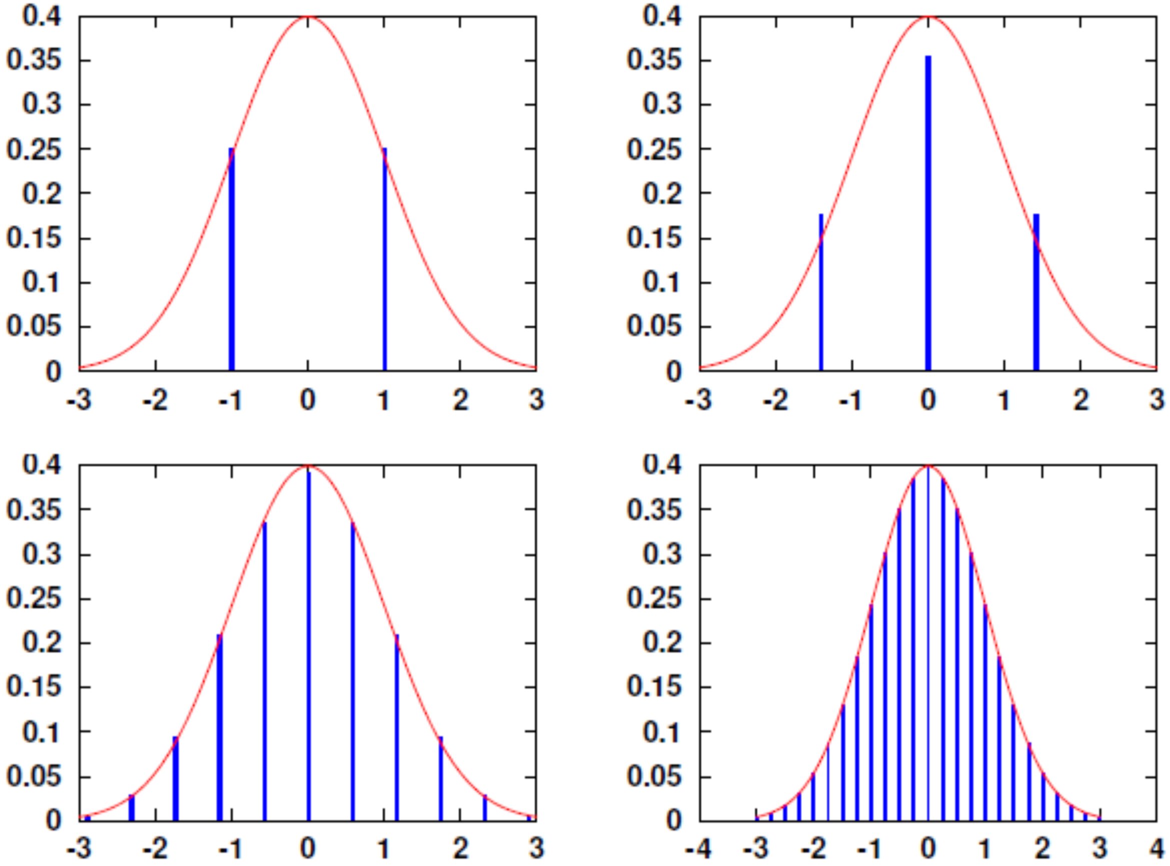
CLT: pictures 2

The standardized average of n i.i.d. exponential random variables with $n = 1, 2, 8, 64$.



CLT: pictures 3

The standardized average of n i.i.d. Bernoulli(0.5) random variables with $n = 1, 2, 12, 64$.



CLT: pictures 4

The (non-standardized) average of n Bernoulli(0.5) random variables, with $n = 4, 12, 64$. (Spikier.)

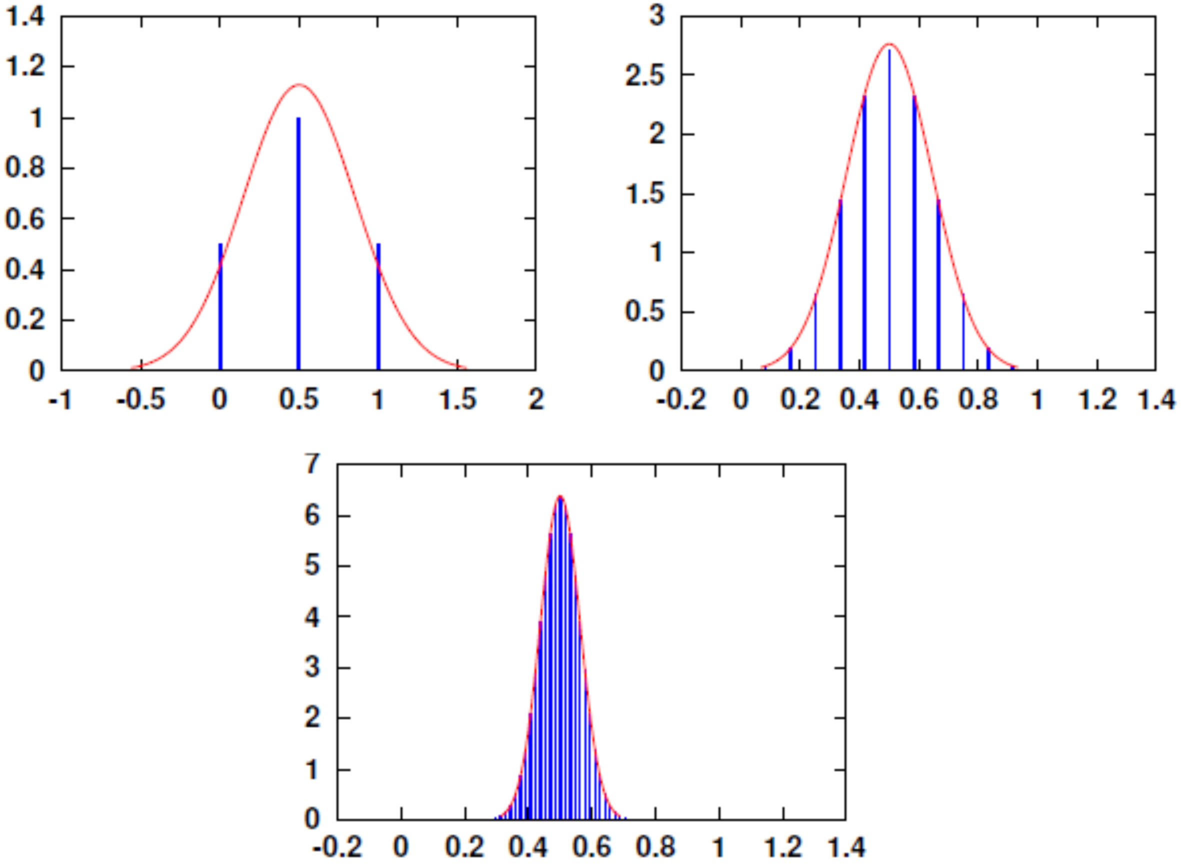


Table Question: Sampling from the standard normal distribution

As a table, produce a single random sample from (an approximate) standard normal distribution.

The table is allowed nine rolls of the 10-sided die.

Note: $\mu = 5.5$ and $\sigma^2 = 8.25$ for a single 10-sided die.

Hint: CLT is about averages.

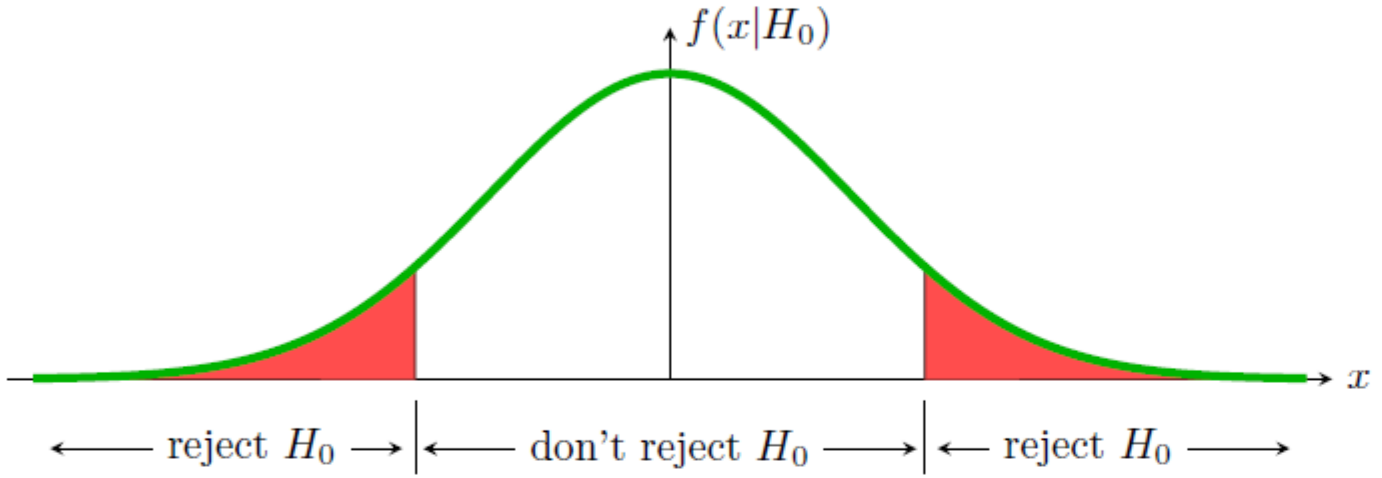
answer: The average of 9 rolls is a sample from the average of 9 independent random variables. The CLT says this average is approximately normal with $\mu = 5.5$ and $\sigma = 8.25/\sqrt{9} = 2.75$
If $\overline{x}$ is the average of 9 rolls then standardizing we get

$$z = \frac{\overline{x} - 5.5}{2.75}$$

is (approximately) a sample from $N(0, 1)$.

Hypothesis testing

Understand this figure



- x = test statistic
- $f(x|H_0)$ = pdf of null distribution = green curve
- Rejection region is a portion of the x-axis.
- Significance = probability over the rejection region = red area.

Simple and composite hypotheses

Simple hypothesis: the sampling distribution is fully specified. Usually the parameter of interest has a specific value.

Composite hypotheses: the sampling distribution is not fully specified. Usually the parameter of interest has a range of values.

Example. A coin has probability θ of heads. Toss it 30 times and let x be the number of heads.

(i) H: $\theta = 0.4$ is simple. $x \sim \text{binomial}(30, 0.4)$.

(ii) H: $\theta > 0.4$ is composite. $x \sim \text{binomial}(30, \theta)$ depends on which value of θ is chosen.

Extreme data and p-values

Hypotheses: H_0, H_A.

Test statistic: value: x, random variable X.

Null distribution: $f(x|H_0)$ (assumes the null hypothesis is true)

Sides: H_A determines if the rejection region is one or two-sided.

Rejection region/Significance: $P(x \text{ in rejection region} \mid H_0) = \alpha$.

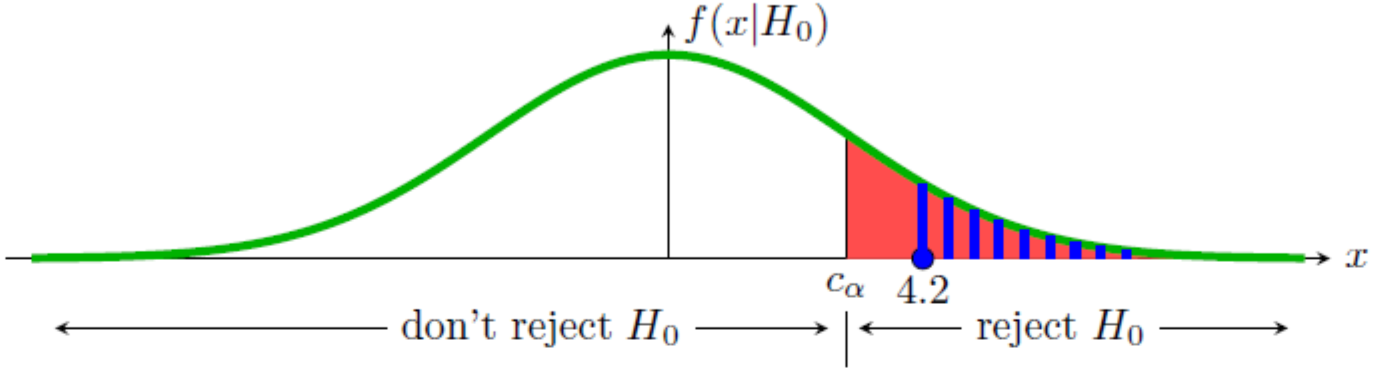
The p-value is a computational tool to check if the test statistic is in the rejection region. It is also a measure of the evidence for rejecting H_0.

p-value: $P(\textbf{data at least as extreme as } x \mid H_0)$

Data at least as extreme: Determined by the sided-ness of the rejection region.

Extreme data and p-values

Example. Suppose we have the right-sided rejection region shown below. Also suppose we see data with test statistic $x = 4.2$. Should we reject H_0?



answer: The test statistic is in the rejection region, so reject H_0.

Alternatively: blue area $<$ red area

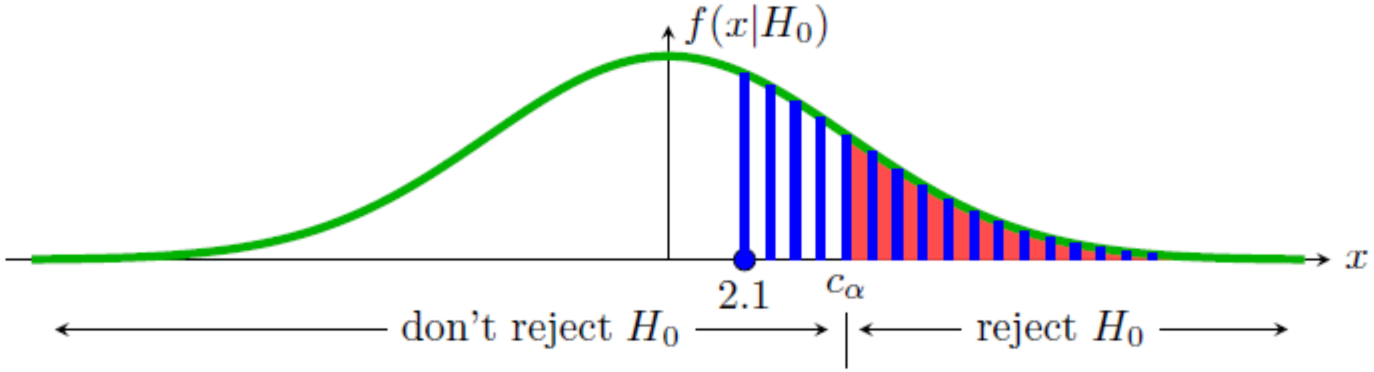
Significance: $\alpha = P(x \text{ in rejection region} \,|\, H_0) =$ red area.

p-value: $p = P(\text{data at least as extreme as } x \,|\, H_0) =$ blue area.

Since, $p < \alpha$ we reject H_0.

Extreme data and p-values

Example. Now suppose $x = 2.1$ as shown. Should we reject H_0?



<u>**answer:**</u> The test statistic is not in the rejection region, so don't reject H_0.

Alternatively: blue area $>$ red area

Significance: $\alpha = P(x \text{ in rejection region} \mid H_0) = \text{red area}$.

p-value: $p = P(\text{data at least as extreme as } x \mid H_0) = \text{blue area}$.

Since, $p > \alpha$ we don't reject H_0.

Critical values

Critical values:

- The boundary of the rejection region are called critical values.
- Critical values are labeled by the probability to their right.
- They are complementary to quantiles: $c_{0.1} = q_{0.9}$

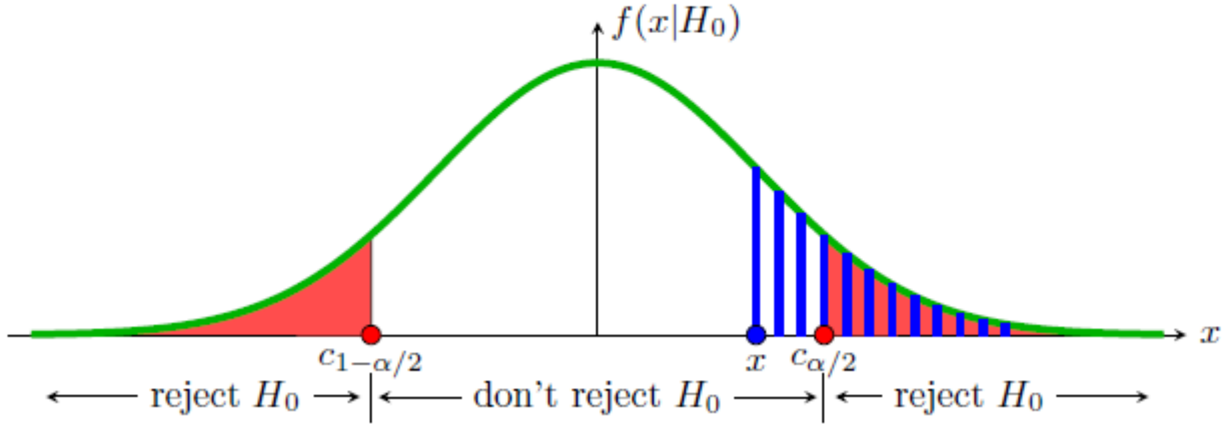
Two-sided p-values

These are trickier: what does 'at least as extreme' mean in this case?

Remember the p-value is a trick for deciding if the test statistic is in the region.

If the significance (rejection) probability is split evenly between the left and right tails then

$$p = 2\min(\text{left tail prob. of } x, \text{ right tail prob. of } x)$$



x is outside the rejection region, so $p > \alpha$: do not reject H_0

Student's T-Test

One-sample t-test

- Data: we assume normal data with both μ and σ unknown:
$$x_1, x_2, \ldots, x_n \sim N(\mu, \sigma^2).$$
- Null hypothesis: $\mu = \mu_0$ for some specific value μ_0.
- Test statistic:
$$t = \frac{\overline{x} - \mu_0}{s/\sqrt{n}}$$
where
$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})^2.$$
Here t is the *Studentized mean* and s^2 is the *sample variance*.
- Null distribution: $f(t \mid H_0)$ is the pdf of $T \sim t(n-1)$, the t distribution with $n-1$ degrees of freedom.

Two-sample t-test: equal variances

Data: we assume normal data with μ_x, μ_y and (same) σ unknown:

$$x_1, \ldots, x_n \sim \mathrm{N}(\mu_x, \sigma^2), \quad y_1, \ldots, y_m \sim \mathrm{N}(\mu_y, \sigma^2)$$

Null hypothesis H_0: $\mu_x = \mu_y$.

Pooled variance: $$s_p^2 = \frac{(n-1)s_x^2 + (m-1)s_y^2}{n+m-2}\left(\frac{1}{n} + \frac{1}{m}\right).$$

Test statistic: $$t = \frac{\bar{x} - \bar{y}}{s_p}$$

Null distribution: $f(t \mid H_0)$ is the pdf of $T \sim t(n+m-2)$

In general (so we can compute power) we have

$$\frac{(\bar{x} - \bar{y}) - (\mu_x - \mu_y)}{s_p} \sim t(n+m-2)$$

Note: there are more general formulas for unequal variances.

Example

$$t = \frac{(\sum D)/N}{\sqrt{\frac{\sum D^2 - \left(\frac{(\sum D)^2}{N}\right)}{(N-1)(N)}}}$$

$$t = \frac{-73/11}{\sqrt{\frac{1131 - \left(\frac{(-73)^2}{11}\right)}{(11-1)(11)}}}$$

$$t = \frac{-73/11}{\sqrt{\frac{1131 - \left(\frac{5329}{11}\right)}{110}}}$$

$$t = -\ 2.74$$

Subject #	Score 1	Score 2	X-Y	(X-Y)^2
1	3	20	-17	289
2	3	13	-10	100
3	3	13	-10	100
4	12	20	-8	64
5	15	29	-14	196
6	16	32	-16	256
7	17	23	-6	36
8	19	20	-1	1
9	23	25	-2	4
10	24	15	9	81
11	32	30	2	4
		SUM:	**-73**	**1131**