Hypothesis testing 1: z-test and t-test

Y. Polyanskiy, D. Shah, J. Tsitsiklis

6.S077

2018

Outline:

- Examples & types
- Relation to philosophy of science and law
- Formal definitions
- **z-test and t-test**
- **The Wald test**

Example of a statistical HT

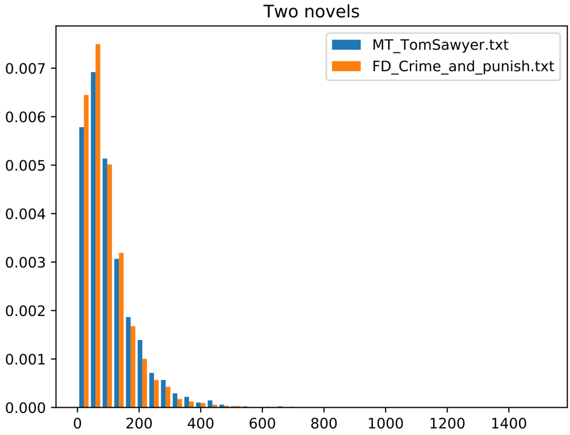
- Hypothesis: F. Dostoevsky's sentences are longer than M. Twain's
- How to confirm/deny it?

Example of a statistical HT

- Hypothesis: F. Dostoevsky's sentences are longer than M. Twain's
- How to confirm/deny it?
- Use data!
- *Crime and Punishment* and *Tom Sawyer*
- Histogram?

Example of a statistical HT

- Hypothesis: F. Dostoevsky's sentences are longer than M. Twain's
- How to confirm/deny it?
- Use data!
- *Crime and Punishment* and *Tom Sawyer*
- Histogram?



Example of a statistical HT

- Hypothesis: F. Dostoevsky's sentences are longer than M. Twain's
- How to confirm/deny it?
- Use data!
- *Crime and Punishment* and *Tom Sawyer*
- Histogram? Inconclusive...
- Sample mean? This lecturer: aka empirical mean

Example of a statistical HT

- Hypothesis: F. Dostoevsky's sentences are longer than M. Twain's
- How to confirm/deny it?
- Use data!
- *Crime and Punishment* and *Tom Sawyer*
- Histogram? Inconclusive...
- Sample mean? This lecturer: aka empirical mean

$$\hat{\mu}_{FD} = 93.8 \text{ char.}, \quad \hat{\mu}_{MT} = 105.6 \text{ char.}$$

- Is difference significant to declare $\mu_{FD} \not> \mu_{MT}$?
  Or could this fluctuation be attributed to chance?

Example of a statistical HT

- Hypothesis: F. Dostoevsky's sentences are longer than M. Twain's
- How to confirm/deny it?
- Use data!
- *Crime and Punishment* and *Tom Sawyer*
- Histogram? Inconclusive...
- Sample mean? This lecturer: aka empirical mean

$$\hat{\mu}_{FD} = 93.8 \text{ char.}, \quad \hat{\mu}_{MT} = 105.6 \text{ char.}$$

- Is difference significant to declare $\mu_{FD} \not> \mu_{MT}$?
  Or could this fluctuation be attributed to chance?
- Main topic of Hypothesis Testing.

Actual real-world examples of statistical HT

More examples:

- Hyp: Is this stock price growing faster than market?
  Data: stock prices
  Type: one-sample test for mean

Actual real-world examples of statistical HT

More examples:

- Hyp: Is this stock price growing faster than market?
  Data: stock prices
  Type: one-sample test for mean
- Hyp: Is this drug better than placebo (nothing)?
  Data: patient health records
  Type: two-sample test for mean difference

Actual real-world examples of statistical HT

More examples:

- Hyp: Is this stock price growing faster than market?
  Data: stock prices
  Type: one-sample test for mean
- Hyp: Is this drug better than placebo (nothing)?
  Data: patient health records
  Type: two-sample test for mean difference
- Hyp: Is NN1 better than NN2 for cat/dog?
  Data: 2x2 table NN1/NN2 vs correct/incorrect
  Type: analysis of contingency tables

Actual real-world examples of statistical HT

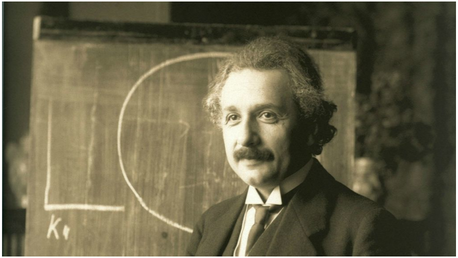
More examples:

- Hyp: Is this stock price growing faster than market?
  Data: stock prices
  Type: one-sample test for mean
- Hyp: Is this drug better than placebo (nothing)?
  Data: patient health records
  Type: two-sample test for mean difference
- Hyp: Is NN1 better than NN2 for cat/dog?
  Data: 2x2 table NN1/NN2 vs correct/incorrect
  Type: analysis of contingency tables

Remarks:

- Hypotheses start life as vague statements
- ... then we add data and probabilistic model
- ... after this hypothesis becomes testable
- Key difference from classification: no prior examples (!) and asymmetric hypotheses (!)

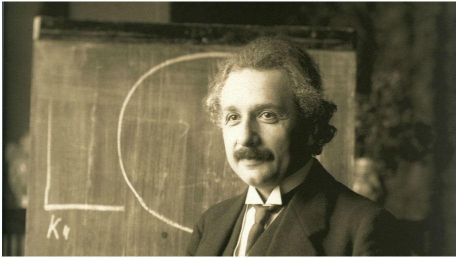
Philosophy of science and HT



"No amount of experimentation can ever prove me right; a single experiment can prove me wrong."

- Cornerstone of scientific method: falsifiability (K. Popper)
- Theory makes predictions. Supported by experiments?
  Then, theory is *valid*
- ... as opposed to **true**.
- Key point: Can never **prove** null (theory). But can REJECT

Philosophy of science and HT



"No amount of experimentation can ever prove me right; a single experiment can prove me wrong."

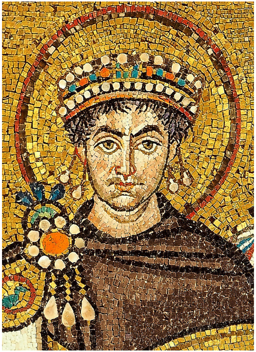
- Cornerstone of scientific method: falsifiability (K. Popper)
- Theory makes predictions. Supported by experiments?
  Then, theory is *valid*
- ... as opposed to **true**.
- Key point: Can never **prove** null (theory). But can REJECT
- Hypotheses are asymmetric!

Law and HT

Ei incumbit probatio qui dicit, non qui negat

–

Proof lies on him who asserts, not on him who denies.



Justinian I

- In law, presumption of innocence
- Null (not guilty) has to be rejected "beyond reasonable doubt".
- ... compare "innocent" vs. "found not guilty"
- ... guilt "proved" by showing that data is very unlikely under null H.
- Equivalences: lenient jury $\iff$ low α

  amount of evidence $\iff$ sample size, etc.

Example of wrong (but "sciency") reasoning (F)

- Consider testing for validity of birth records of a city.
- First consider this record:

$$M, M, M, F, F, F, F, M, M, M, F, F, F, F, M, F, M, F, F, M,$$
$$M, M, F, M, F, F, F, M, F, M, F, M, F, M, M, F, F, F, M, M,$$
$$F, F, M, F, M, M, F, F, F, F, F, F, M, M, M, M, F, F, M, M,$$
$$M, F, F, M, F, F, M, F, F, F, M, M, F, M, M, M, F, M, M, M,$$
$$M, F, M, M, M, M, F, F, M, F, F, M, M, M, F, M, F, M, F, M$$

- Looks ok...

Example of wrong (but "sciency") reasoning (F)

- Consider testing for validity of birth records of a city.
- First consider this record:

$$M, M, M, F, F, F, F, M, M, M, F, F, F, F, M, F, M, F, F, M,$$
$$M, M, F, M, F, F, F, M, F, M, F, M, F, M, M, F, F, F, M, M,$$
$$F, F, M, F, M, M, F, F, F, F, F, F, M, M, M, M, F, F, M, M,$$
$$M, F, F, M, F, F, M, F, F, F, M, M, F, M, M, M, F, M, M, M,$$
$$M, F, M, M, M, M, F, F, M, F, F, M, M, M, F, M, F, M, F, M$$

- Looks ok...
- Second, consider this record:

$$M, M, M, M, M, M, M, M, M, M, M, M, M, M, M, M, M, M,$$
$$M, M, M, M, M, M, M, M, M, M, M, M, M, M, M, M, M, M,$$
$$M, M, M, M, M, M, M, M, M, M, M, M, M, M, M, M, M, M,$$
$$M, M, M, M, M, M, M, M, M, M, M, M, M, M, M, M, M, M,$$
$$M, M, M, M, M, M, M, M, M, M, M, M, M, M, M, M, M, M,$$
$$M, M, M, M, M, M, M, M, M, M$$

- Hmm... Is it valid?

Example of wrong (but "sciency") reasoning (F)

- Consider testing for validity of birth records of a city.
- First consider this record:

  M, M, M, F, F, F, F, M, M, M, F, F, F, F, M, F, M, F, F, M,
  M, M, F, M, F, F, F, M, F, M, F, M, F, M, M, F, F, F, M, M,
  F, F, M, F, M, M, F, F, F, F, F, F, M, M, M, M, F, F, M, M,
  M, F, F, M, F, F, M, F, F, F, M, M, F, M, M, M, F, M, M, M,
  M, F, M, M, M, M, F, F, M, F, F, M, M, M, F, M, F, M, F, M

- Looks ok...
- Second, consider this record:

  M, M, M, M, M, M, M, M, M, M, M, M, M, M, M, M, M, M,
  M, M, M, M, M, M, M, M, M, M, M, M, M, M, M, M, M, M,
  M, M, M, M, M, M, M, M, M, M, M, M, M, M, M, M, M, M,
  M, M, M, M, M, M, M, M, M, M, M, M, M, M, M, M, M, M,
  M, M, M, M, M, M, M, M, M, M, M, M, M, M, M, M, M, M,
  M, M, M, M, M, M, M, M, M, M

- Hmm... Is it valid?
- In nature, $\mathbb{P}[M] = 0.51$. Compare probabilities of two records.

Example of wrong (but “sciency”) reasoning (F)

- Consider testing for validity of birth records of a city.
- First consider this record:

$$M, M, M, F, F, F, F, M, M, M, F, F, F, F, M, F, M, F, F, M,$$
$$M, M, F, M, F, F, F, M, F, M, F, M, F, M, M, F, F, F, M, M,$$
$$F, F, M, F, M, M, F, F, F, F, F, F, M, M, M, M, F, F, M, M,$$
$$M, F, F, M, F, F, M, F, F, F, M, M, F, M, M, M, F, M, M, M,$$
$$M, F, M, M, M, M, F, F, M, F, F, M, M, M, F, M, F, M, F, M$$

- Looks ok...
- Second, consider this record:

$$M, M, M, M, M, M, M, M, M, M, M, M, M, M, M, M, M, M,$$
$$M, M, M, M, M, M, M, M, M, M, M, M, M, M, M, M, M, M,$$
$$M, M, M, M, M, M, M, M, M, M, M, M, M, M, M, M, M, M,$$
$$M, M, M, M, M, M, M, M, M, M, M, M, M, M, M, M, M, M,$$
$$M, M, M, M, M, M, M, M, M, M, M, M, M, M, M, M, M, M,$$
$$M, M, M, M, M, M, M, M, M, M$$

- Hmm... Is it valid?
- In nature, $\mathbb{P}[M] = 0.51$. Compare probabilities of two records.
- Second record is even more likely!

Formal setting for HT

Definition

Statistical hypotheses:

- H : data $X_1, \ldots, X_n$ distributed according to $P \in \mathcal{C}_0$
- K : data $X_1, \ldots, X_n$ distributed according to $P \in \mathcal{C}_1$

where $\mathcal{C}_0, \mathcal{C}_1$ are COLLECTIONS OF DISTRIBUTIONS.

Formal setting for HT

Definition

Statistical hypotheses:

- H : data $X_1, \ldots, X_n$ distributed according to $P \in \mathcal{C}_0$
- K : data $X_1, \ldots, X_n$ distributed according to $P \in \mathcal{C}_1$

where $\mathcal{C}_0, \mathcal{C}_1$ are COLLECTIONS OF DISTRIBUTIONS.

Remarks:

- **ALWAYS (!)** make sure you can formulate in the form above
- Most common special case: X_i are i.i.d. from P
- ... So will just write things like:

$$H : \mathbb{E}[X] = 0 \qquad \text{vs.} \qquad K : \mathbb{E}[X] \neq 0 \,.$$

- Caution: for subsets of population (e.g. polls), iid holds only approximately
  Rule of thumb: iid if sample size is $< 10\%$ of population
- Example: 10 out of 30 is not iid, 1000 out of $7 \cdot 10^9$ is iid.

Formal setting for HT

Definition

Statistical hypotheses:

- H : data $X_1, \dots, X_n$ distributed according to $P \in \mathcal{C}_0$
- K : data $X_1, \dots, X_n$ distributed according to $P \in \mathcal{C}_1$

where $\mathcal{C}_0, \mathcal{C}_1$ are COLLECTIONS OF DISTRIBUTIONS.

Remarks:

- **ALWAYS (!)** make sure you can formulate in the form above
- Most common special case: X_i are i.i.d. from P
- ... S

AGAIN: hypothesis = collection of distributions

- Caution: for subsets of population (e.g. polls), iid holds only approximately
  Rule of thumb: iid if sample size is $< 10\%$ of population
- Example: 10 out of 30 is not iid, 1000 out of $7 \cdot 10^9$ is iid.

Roadmap

Tests we will learn:

- One-sample tests:
  1. for mean of population: $\mathbb{E}[X] = \mu_0$ vs $\mathbb{E}[X] \neq \mu_0$
  2. for other parameters: $\theta \in \Theta_0$ vs $\theta \notin \Theta_0$
  3. generalized likelihood-ratio test: $X \sim$ Uniform vs $X \sim$ not Uniform
  4. testing normality: $X \sim \mathcal{N}(0,1)$ vs $X \nsim \mathcal{N}(0,1)$
- Two-sample tests:
  1. Equality of means: $\mathbb{E}[X] = \mathbb{E}[Y]$ vs. $\mathbb{E}[X] \neq \mathbb{E}[Y]$
  2. Equality of distributions: $P_X = P_Y$ vs. $P_X \neq P_Y$
  3. Testing independence: $X \perp\!\!\!\perp Y$ vs $X \not\perp\!\!\!\perp Y$

$\boxed{z\text{-test and } t\text{-test}}$

One-sample tests for mean: common sense

- Setting:
  - iid samples $X_1, \ldots, X_n$ are observed
  - Goal: test if $\mathbb{E}[X] = 0$ or $\mathbb{E}[X] \neq 0$

One-sample tests for mean: common sense

- Setting:
  - iid samples $X_1, \ldots, X_n$ are observed
  - Goal: test if $\mathbb{E}[X] = 0$ or $\mathbb{E}[X] \neq 0$
- Bro-data-science method:
  - Take first $n/10$ samples, compute empirical mean. Say you got 0.1
  - Take next $n/10$ samples, empirical mean 0.09
  - ... keep going, get sample means like:

$$0.1, 0.09, 0.11, 0.1, \cdots$$

  - Conclude, true mean is about 0.1 ± 0.01
  - Move on to REJECT null.
- Is this the end of our 5 lectures?

One-sample tests for mean: common sense

- Setting:
  - iid samples $X_1, \ldots, X_n$ are observed
  - Goal: test if $\mathbb{E}[X] = 0$ or $\mathbb{E}[X] \neq 0$
- Bro-data-science method:
  - Take first $n/10$ samples, compute empirical mean. Say you got 0.1
  - Take next $n/10$ samples, empirical mean 0.09
  - ... keep going, get sample means like:

$$0.1, 0.09, 0.11, 0.1, \cdots$$

  - Conclude, true mean is about 0.1 ± 0.01
  - Move on to REJECT null.
- Is this the end of our 5 lectures?
- Well, this method is bad because (at least)
  - Very wasteful of data (will need huge n to work)
  - Does not provide clear guarantees (even asymptotic in n)
  - Not well-specified (what if means were $0.01, 0.001$?)

One-sample tests for mean: machine learning (F)

- Setting:
  - iid samples $X_1, \ldots, X_n$ are observed
  - Goal: test if $\mathbb{E}[X] = 0$ or $\mathbb{E}[X] \neq 0$
- ML method:
  - Generate 10^9 datasets from $\mu = 0$
  - Generate 10^9 datasets from $\mu \neq 0$
  - Train classifier (DNN, of course)
  - Ask it to classify the observed example
- What is wrong with this?

One-sample tests for mean: machine learning (F)

- Setting:
  - iid samples $X_1, \ldots, X_n$ are observed
  - Goal: test if $\mathbb{E}[X] = 0$ or $\mathbb{E}[X] \neq 0$
- ML method:
  - Generate 10^9 datasets from $\mu = 0$
  - Generate 10^9 datasets from $\mu \neq 0$
  - Train classifier (DNN, of course)
  - Ask it to classify the observed example
- What is wrong with this?
  - Non-reproducible (same data leads to different answers)
  - What distributions to generate from?
  - No false-positive guarantees

One-sample tests for mean: machine learning (F)

- Setting:
  - iid samples $X_1, \ldots, X_n$ are observed
  - Goal: test if $\mathbb{E}[X] = 0$ or $\mathbb{E}[X] \neq 0$
- ML method:
  - Generate 10^9 datasets from $\mu = 0$
  - Generate 10^9 datasets from $\mu \neq 0$
  - Train classifier (DNN, of course)
  - Ask it to classify the observed example
- What is wrong with this?
  - Non-reproducible (same data leads to different answers)
  - What distributions to generate from?
  - No false-positive guarantees
- How binary HT is different from 'cat/dog' classification:
  - Hypotheses are not symmetric
  - No good choice for Bayesian priors on P's in $\mathcal{C}_0$ or $\mathcal{C}_1$
  - Futuristic: train on actually labeled experiments?..

One-sample tests for mean: z-test

Let $\mu \triangleq \mathbb{E}[X]$ and consider one of

1. $\mu = \mu_0$ vs. $\mu \neq \mu_0$
2. $\mu = \mu_0$ vs. $\mu > \mu_0$
3. $\mu < \mu_0$ vs. $\mu > \mu_0$

One-sample tests for mean: z-test

Let $\mu \triangleq \mathbb{E}[X]$ and consider one of

$$\left.\begin{array}{ll} \text{1. } \mu = \mu_0 \text{ vs. } \mu \neq \mu_0 & \text{"two-sided"} \end{array}\right.$$

$$\left.\begin{array}{l} \text{2. } \mu = \mu_0 \text{ vs. } \mu > \mu_0 \\ \text{3. } \mu < \mu_0 \text{ vs. } \mu > \mu_0 \end{array}\right\} \text{"one-sided"}$$

One-sample tests for mean: z-test

Let $\mu \triangleq \mathbb{E}[X]$ and consider one of

1. $\mu = \mu_0$ vs. $\mu \neq \mu_0$ "two-sided"
2. $\mu = \mu_0$ vs. $\mu > \mu_0$
3. $\mu < \mu_0$ vs. $\mu > \mu_0$

} "one-sided"

"Known variance" special case:

- Suppose under H: $\mathrm{Var}[X] = \sigma_0^2$

z-statistic

$$Z \triangleq \frac{\sum_{i=1}^{n}(X_i - \mu_0)}{\sqrt{n\sigma_0^2}}$$

- Depending on the version of HTs the z-test is
  1. If $|z| > t$ then REJECT null
  2. If $z > t_1$ the REJECT null
  3. ... same test ...

One-sample tests for mean: z-test

Let $\mu \triangleq \mathbb{E}[X]$ and consider one of

1. $\mu = \mu_0$ vs. $\mu \neq \mu_0$ — "two-sided"
2. $\mu = \mu_0$ vs. $\mu > \mu_0$
3. $\mu < \mu_0$ vs. $\mu > \mu_0$

$\left.\begin{array}{l} 2 \\ 3 \end{array}\right\}$ "one-sided"

"Known variance" special case:

- Suppose under H: $\mathrm{Var}[X] = \sigma_0^2$

z-statistic

$$Z \triangleq \frac{\sum_{i=1}^{n}(X_i - \mu_0)}{\sqrt{n\sigma_0^2}} \quad = \quad (\hat{\mu} - \mu_0)\sqrt{\frac{n}{\sigma_0^2}}$$

- Depending on the version of HTs the z-test is
  1. If $|z| > t$ then REJECT null
  2. If $z > t_1$ the REJECT null
  3. . . . same test . . .

Understanding z-test

Carefully restating the null-hypothesis in case ❶:

- X_i are iid from some P
- P has mean μ_0
- P has variance σ_0^2

Reasoning for z-test:

- From CLT:
$$Z \triangleq \frac{\sum_{i=1}^{n}(X_i - \mu_0)}{\sqrt{n\sigma_0^2}} \overset{d}{\approx} \mathcal{N}(0,1)\,.$$
- Great: Under null z should be very close to standard normal.
- Even better: if $\mu \neq \mu_0$ we will have: $|z| \to \infty$ (with speed of $\sqrt{n}(\mu - \mu_0)$)
- If z looks suspiciously large for a $\mathcal{N}(0,1)$, reject null!

Understanding z-test: selecting threshold

Carefully restating the null-hypothesis in case ❶:

- X_i are iid from some P
- P has mean μ_0
- P has variance σ_0^2

Reasoning for z-test:

- From CLT:

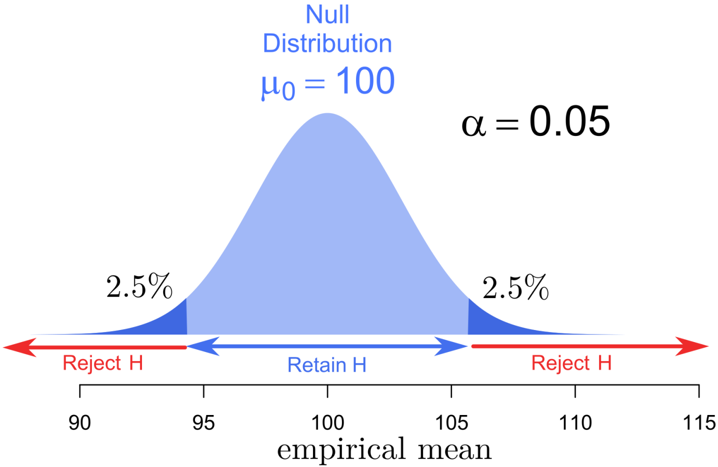
$$Z \triangleq \frac{\sum_{i=1}^{n}(X_i - \mu_0)}{\sqrt{n\sigma_0^2}} \overset{d}{\approx} \mathcal{N}(0, 1) \,.$$

- Q: What Z is suspiciously large?
- A: Depends on required probability of false-positive!
  (... Yes, HT is always subject to error probability.)
- Under null

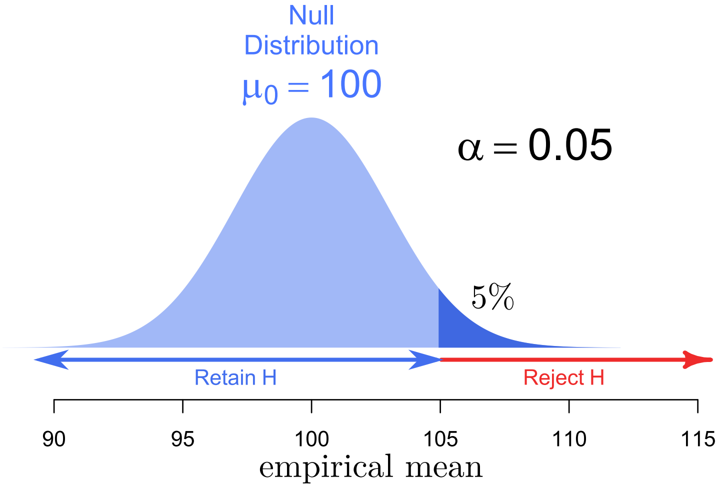
$$\mathbb{P}[|Z| > t] \approx 2 \times \texttt{scipy.stats.norm.sf(t)}$$

- Nicknames for false-positive rate:
  size of test, significance level, type-1 error, ~~p-value~~

Thresholding Z: two-tailed case



Thresholding Z: one-tailed case



Seattle

- Famous for its rain
- Naive probability: $\mathbb{P}[\text{rain}] = 1/2$?

Example of using z-test

$$H: \mathbb{P}[\textit{Rain in Seattle}] = 1/2$$

- $\mu_0 = 1/2$, $\sigma_0^2 = 1/4$. Test of size $\alpha = 0.05$ will be:

REJECT null if $|Z| > 1.96$

Example of using z-test

$$H: \mathbb{P}[\textit{Rain in Seattle}] = 1/2$$

- $\mu_0 = 1/2$, $\sigma_0^2 = 1/4$. Test of size $\alpha = 0.05$ will be:

REJECT null if $|Z| > 1.96$

- Take random n days from Seattle weather data

DATE	RAIN
1950-05-25	False
1957-07-14	True
1960-05-18	False
1982-02-10	False
1994-08-11	False
1994-09-24	False
1996-07-29	False
1997-06-01	True
1997-12-22	True
2001-02-14	False

$Z = -1.26$

Example of using z-test

$$H: \mathbb{P}[\textit{Rain in Seattle}] = 1/2$$

- $\mu_0 = 1/2$, $\sigma_0^2 = 1/4$. Test of size $\alpha = 0.05$ will be:

REJECT null if $|Z| > 1.96$

- Take random n days from Seattle weather data
- $n = 10$ samples, got Z-scores:

$$Z = 0.0, -1.89, -0.63, 0.63, -0.63$$

Test: accept null

DATE	RAIN
1950-05-25	False
1957-07-14	True
1960-05-18	False
1982-02-10	False
1994-08-11	False
1994-09-24	False
1996-07-29	False
1997-06-01	True
1997-12-22	True
2001-02-14	False

$Z = -1.26$

Example of using z-test

$$H: \mathbb{P}[\textit{Rain in Seattle}] = 1/2$$

- $\mu_0 = 1/2$, $\sigma_0^2 = 1/4$. Test of size $\alpha = 0.05$ will be:

REJECT null if $|Z| > 1.96$

- Take random n days from Seattle weather data
- $n = 10$ samples, got Z-scores:

$$Z = 0.0, -1.89, -0.63, 0.63, -0.63$$

Test: accept null

- $n = 100$, got Z-scores:

$$Z = -3.2, -2.0, -1.4, -3.6, -1.2$$

Test: sometimes reject sometimes accept

DATE	RAIN
1950-05-25	False
1957-07-14	True
1960-05-18	False
1982-02-10	False
1994-08-11	False
1994-09-24	False
1996-07-29	False
1997-06-01	True
1997-12-22	True
2001-02-14	False

$$Z = -1.26$$

Example of using z-test

$$H: \mathbb{P}[\textit{Rain in Seattle}] = 1/2$$

- $\mu_0 = 1/2$, $\sigma_0^2 = 1/4$. Test of size $\alpha = 0.05$ will be:

REJECT null if $|Z| > 1.96$

- Take random n days from Seattle weather data
- $n = 10$ samples, got Z-scores:

$$Z = 0.0, -1.89, -0.63, 0.63, -0.63$$

Test: accept null

- $n = 100$, got Z-scores:

$$Z = -3.2, -2.0, -1.4, -3.6, -1.2$$

Test: sometimes reject sometimes accept

- $n = 1000$, got Z-scores:

$$Z = -5.76, -2.72, -3.29, -3.92, -4.42$$

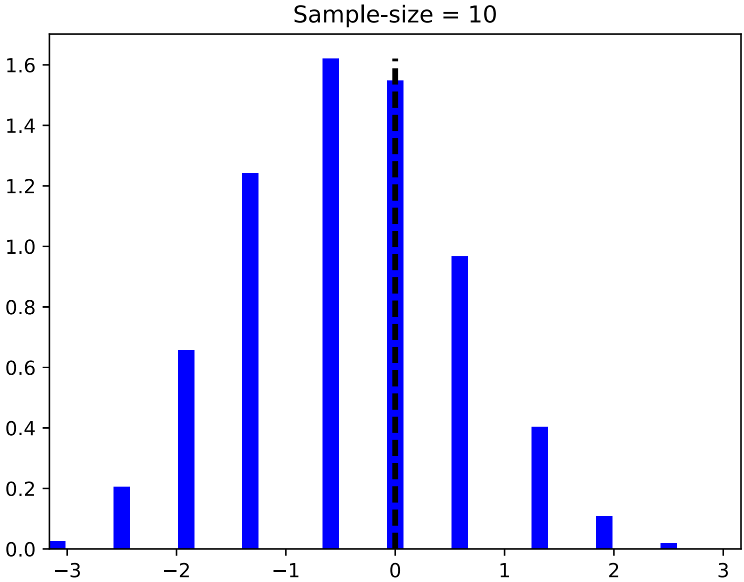
Test: reject null (always)

(FYI: over 1948-2017 $\hat{P}$[rain] $= 42.6\%$)

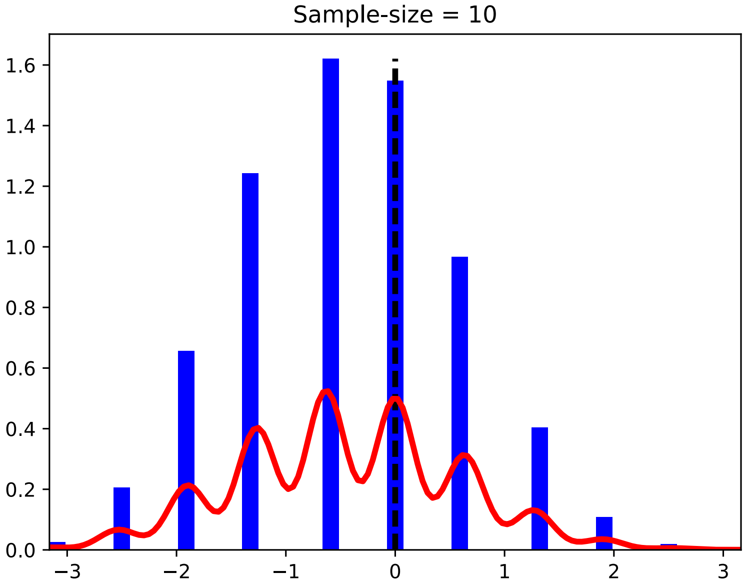
DATE	RAIN
1950-05-25	False
1957-07-14	True
1960-05-18	False
1982-02-10	False
1994-08-11	False
1994-09-24	False
1996-07-29	False
1997-06-01	True
1997-12-22	True
2001-02-14	False

$$Z = -1.26$$

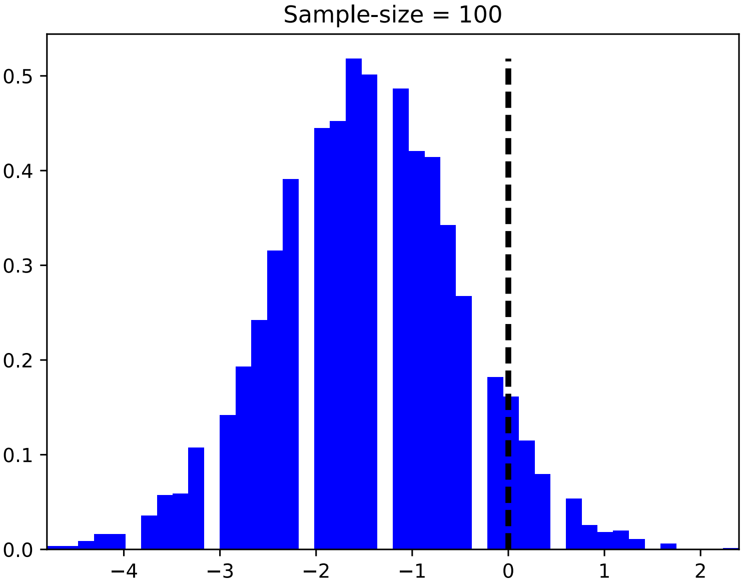
Histogram of z-scores: effect of sample size



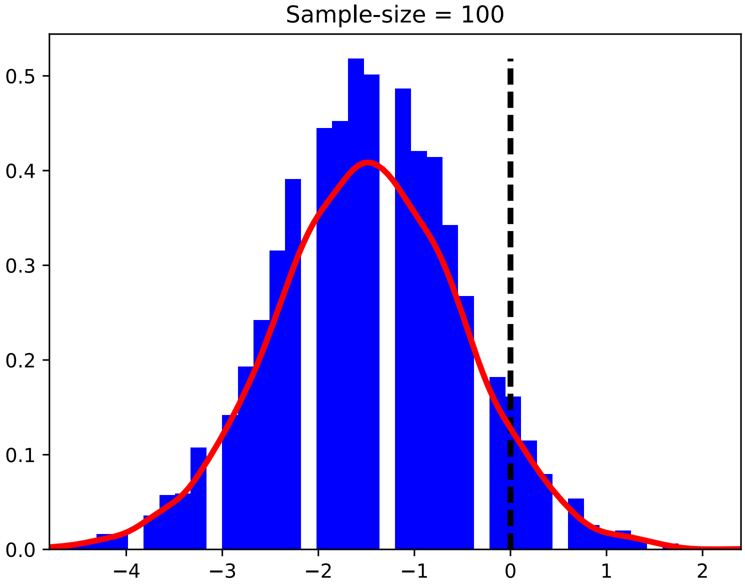
Histogram of z-scores: effect of sample size



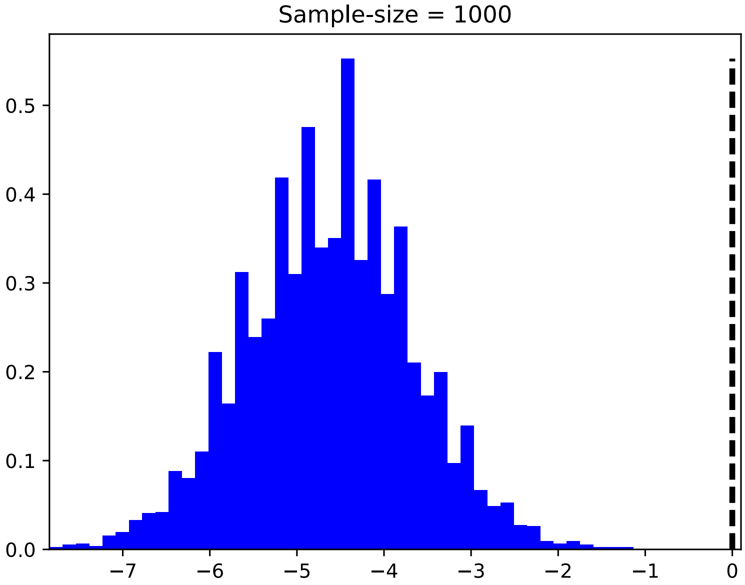
Histogram of z-scores: effect of sample size



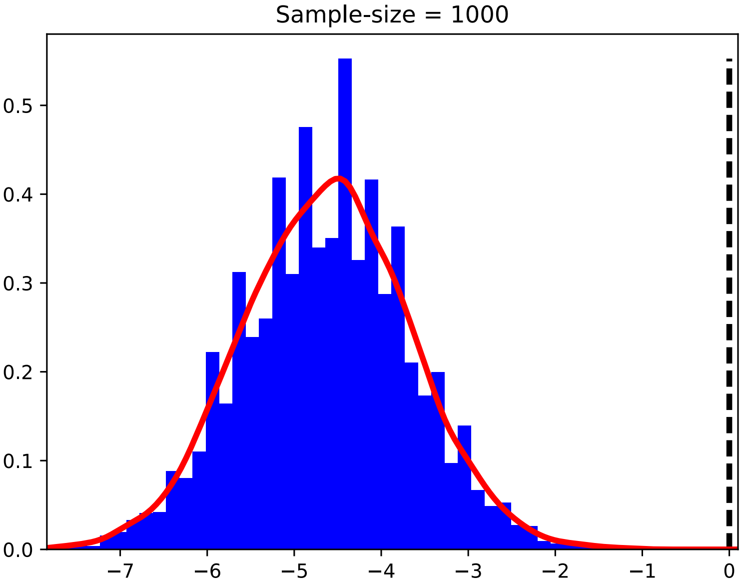
Histogram of z-scores: effect of sample size



Histogram of z-scores: effect of sample size



Histogram of z-scores: effect of sample size



One-sample tests for mean: t-test

Let $\mu \triangleq \mathbb{E}[X]$ and consider one of

1. $\mu = \mu_0$ vs. $\mu \neq \mu_0$ "two-sided"
2. $\mu = \mu_0$ vs. $\mu > \mu_0$
3. $\mu < \mu_0$ vs. $\mu > \mu_0$

$\left.\begin{array}{l}2.\\3.\end{array}\right\}$ "one-sided"

- **"Known variance" case:** use z-statistic $Z = (\hat{\mu} - \mu_0)\sqrt{\frac{n}{\sigma_0^2}}$

One-sample tests for mean: t-test

Let $\mu \triangleq \mathbb{E}[X]$ and consider one of

1. $\mu = \mu_0$ vs. $\mu \neq \mu_0$ "two-sided"
2. $\mu = \mu_0$ vs. $\mu > \mu_0$
3. $\mu < \mu_0$ vs. $\mu > \mu_0$

$\left.\begin{array}{l} 2. \\ 3. \end{array}\right\}$ "one-sided"

- **"Known variance" case:** use z-statistic $Z = (\hat{\mu} - \mu_0)\sqrt{\frac{n}{\sigma_0^2}}$
- **"Unknown variance" case:**

One-sample tests for mean: t-test

Let $\mu \triangleq \mathbb{E}[X]$ and consider one of

1. $\mu = \mu_0$ vs. $\mu \neq \mu_0$ "two-sided"
2. $\mu = \mu_0$ vs. $\mu > \mu_0$
3. $\mu < \mu_0$ vs. $\mu > \mu_0$

$\Big\}$ "one-sided"

- **"Known variance" case:** use z-statistic $Z = (\hat{\mu} - \mu_0)\sqrt{\frac{n}{\sigma_0^2}}$
- **"Unknown variance" case:**

t-statistic

$$T \triangleq (\hat{\mu} - \mu_0)\sqrt{\frac{n}{\widehat{\sigma^2}}}$$

with $\hat{\mu} = \frac{1}{n}\sum_i X_i, \quad \widehat{\sigma^2} = \frac{1}{n-1}\sum_i (X_i - \hat{\mu})^2$

- The resulting t-test: If $|T| \geq t_\alpha$ then REJECT
- t-test is a workhorse of data science
- The heart of HT: "measure effect and normalize by typical deviation"

One-sample tests for mean: t-test

Let $\mu \triangleq \mathbb{E}[X]$ and consider one of

1. $\mu = \mu_0$ vs. $\mu \neq \mu_0$ "two-sided"

- **"Unknown variance" case:**

t-statistic

$$T \triangleq (\hat{\mu} - \mu_0)\sqrt{\frac{n}{\widehat{\sigma^2}}}$$

with $\hat{\mu} = \frac{1}{n}\sum_i X_i, \quad \widehat{\sigma^2} = \frac{1}{n-1}\sum_i (X_i - \hat{\mu})^2$

- How to threshold T?

One-sample tests for mean: t-test

Let $\mu \triangleq \mathbb{E}[X]$ and consider one of

1. $\mu = \mu_0$ vs. $\mu \neq \mu_0$ "two-sided"

- **"Unknown variance" case:**

t-statistic

$$T \triangleq (\hat{\mu} - \mu_0)\sqrt{\frac{n}{\widehat{\sigma^2}}}$$

with $\hat{\mu} = \frac{1}{n}\sum_i X_i$, $\widehat{\sigma^2} = \frac{1}{n-1}\sum_i (X_i - \hat{\mu})^2$

- How to threshold T?
- Under H: $\mathbb{E}[X] = \mu_0, \mathrm{Var}[X] = ???$
- Magic 1: dist. of T does not depend on μ_0 or $\mathrm{Var}[X]$ (PSet 2)

One-sample tests for mean: t-test

Let $\mu \triangleq \mathbb{E}[X]$ and consider one of

1. $\mu = \mu_0$ vs. $\mu \neq \mu_0$ "two-sided"

- **"Unknown variance" case:**

t-statistic

$$T \triangleq (\hat{\mu} - \mu_0)\sqrt{\frac{n}{\widehat{\sigma^2}}}$$

with $\hat{\mu} = \frac{1}{n}\sum_i X_i, \quad \widehat{\sigma^2} = \frac{1}{n-1}\sum_i (X_i - \hat{\mu})^2$

- How to threshold T?
- Under H: $\mathbb{E}[X] = \mu_0, \mathrm{Var}[X] = ???$
- Magic 1: dist. of T does not depend on μ_0 or $\mathrm{Var}[X]$ (PSet 2)
- Magic 2: CLT and LLN (!)
- By LLN $\hat{\mu} \to \mu_0$, $\widehat{\sigma^2} \to \mathrm{Var}[X]$

One-sample tests for mean: t-test

Let $\mu \triangleq \mathbb{E}[X]$ and consider one of

1. $\mu = \mu_0$ vs. $\mu \neq \mu_0$ "two-sided"

- **"Unknown variance" case:**

t-statistic

$$T \triangleq (\hat{\mu} - \mu_0)\sqrt{\frac{n}{\widehat{\sigma^2}}}$$

with $\hat{\mu} = \frac{1}{n}\sum_i X_i, \quad \widehat{\sigma^2} = \frac{1}{n-1}\sum_i (X_i - \hat{\mu})^2$

- How to threshold T?
- Under H: $\mathbb{E}[X] = \mu_0, \mathrm{Var}[X] = ???$
- Magic 1: dist. of T does not depend on μ_0 or $\mathrm{Var}[X]$ (PSet 2)
- Magic 2: CLT and LLN (!)
- By LLN $\hat{\mu} \to \mu_0$, $\widehat{\sigma^2} \to \mathrm{Var}[X]$
- So by CLT: Distribution of $T \to \mathcal{N}(0,1)$ as $n \to \infty$ (iff $\mathrm{Var}[X] < \infty$)

One-sample tests for mean: t-test

Let $\mu \triangleq \mathbb{E}[X]$ and consider one of

1. $\mu = \mu_0$ vs. $\mu \neq \mu_0$ "two-sided"

- **"Unknown variance" case:**

t-statistic

$$T \triangleq (\hat{\mu} - \mu_0)\sqrt{\frac{n}{\widehat{\sigma^2}}}$$

with $\hat{\mu} = \frac{1}{n}\sum_i X_i, \quad \widehat{\sigma^2} = \frac{1}{n-1}\sum_i (X_i - \hat{\mu})^2$

- How to threshold T?
- Under H: $\mathbb{E}[X] = \mu_0, \mathrm{Var}[X] =???$
- Magic 1: dist. of T does not depend on μ_0 or $\mathrm{Var}[X]$ (PSet 2)
- Magic 2: CLT and LLN (!)
- By LLN $\hat{\mu} \to \mu_0$, $\widehat{\sigma^2} \to \mathrm{Var}[X]$
- So by CLT: Distribution of $T \to \mathcal{N}(0,1)$ as $n \to \infty$ (iff $\mathrm{Var}[X] < \infty$)
- Punchline: $\boxed{T \approx \mathcal{N}(0,1) \quad \text{for large } n}$

One-sample tests for mean: t-test

Let $\mu \triangleq \mathbb{E}[X]$ and consider one of

1. $\mu = \mu_0$ vs. $\mu \neq \mu_0$ "two-sided"

- **"Unknown variance" case:**

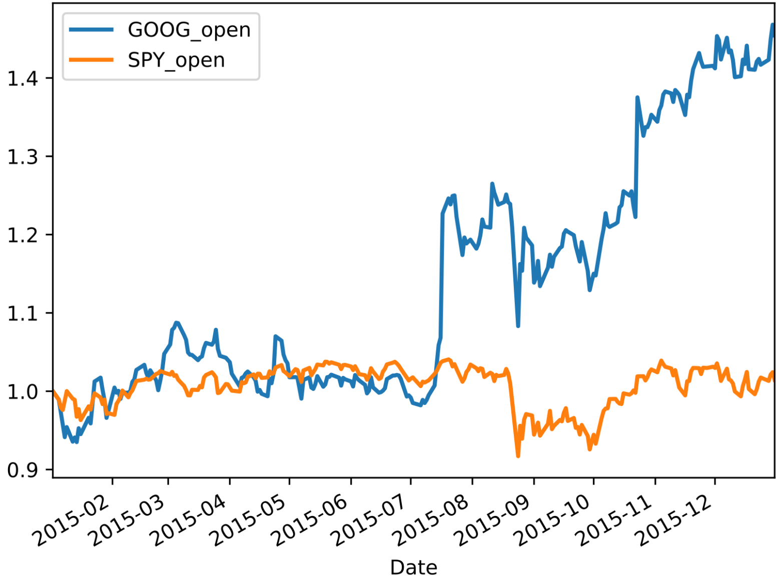
t-statistic

$$T \triangleq (\hat{\mu} - \mu_0)\sqrt{\frac{n}{\widehat{\sigma^2}}}$$

with $\hat{\mu} = \frac{1}{n}\sum_i X_i, \quad \widehat{\sigma^2} = \frac{1}{n-1}\sum_i (X_i - \hat{\mu})^2$

- How to threshold T?
- Under H: $\mathbb{E}[X] = \mu_0, \mathrm{Var}[X] = ???$
- Magic 1: dist. of T does not depend on μ_0 or $\mathrm{Var}[X]$ (PSet 2)
- Magic 2: CLT and LLN (!)
- By LLN $\hat{\mu} \to \mu_0$, $\widehat{\sigma^2} \to \mathrm{Var}[X]$
- So by CLT: Distribution of $T \to \mathcal{N}(0,1)$ as $n \to \infty$ (iff $\mathrm{Var}[X] < \infty$)
- Punchline: $\boxed{T \approx \mathcal{N}(0,1) \quad \text{for large } n}$
- (!) When $X_j \approx$ Gaussian, then $\boxed{T \approx \texttt{sp.stats.t.pdf}(\cdot, \texttt{n} - 1)}$

Google vs S&P500 in 2015



Google vs S&P500 in 2015

In 2015:

- S&P 500 grew by 1.4%
- Google grew by 45.5%
- Is this statistically significant?
- Let's find out...

Google vs S&P500 in 2015

In 2015:

- S&P 500 grew by 1.4%
- Google grew by 45.5%
- Is this statistically significant?
- Let's find out...

Hypothesis testing setup

- Let G_t = price of GOOG, S_t = price of SPY.
- Define $\Delta_t \triangleq \log \frac{G_t}{G_{t-1}} - \log \frac{S_t}{S_{t-1}}$
- ... equals relative daily growth of Google stock w.r.t. S&P500
- Null hypothesis: $H : \mathbb{E}[\Delta] = 0$. Alternative $K : \mathbb{E}[\Delta] \neq 0$

Google vs S&P500 in 2015

In 2015:

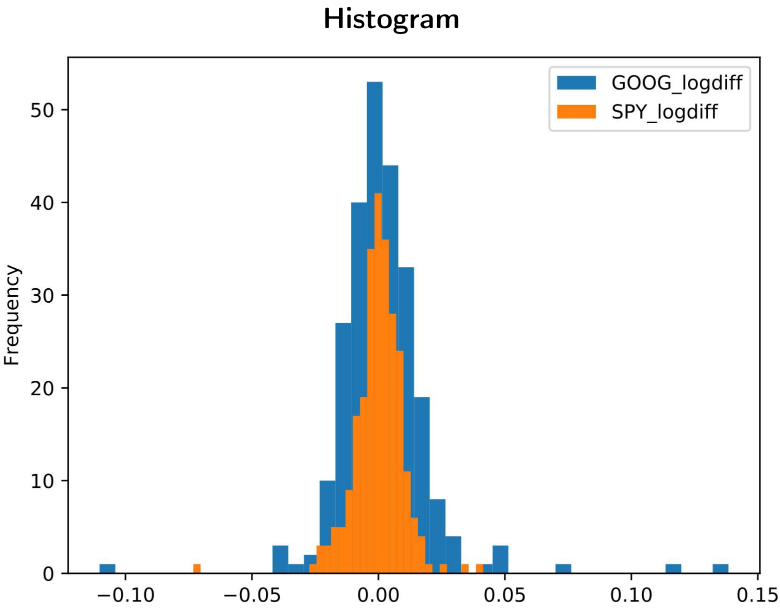
- S&P 500 grew by 1.4%
- Google grew by 45.5%
- Is this statistically significant?
- Let's find out...

Hypothesis testing setup

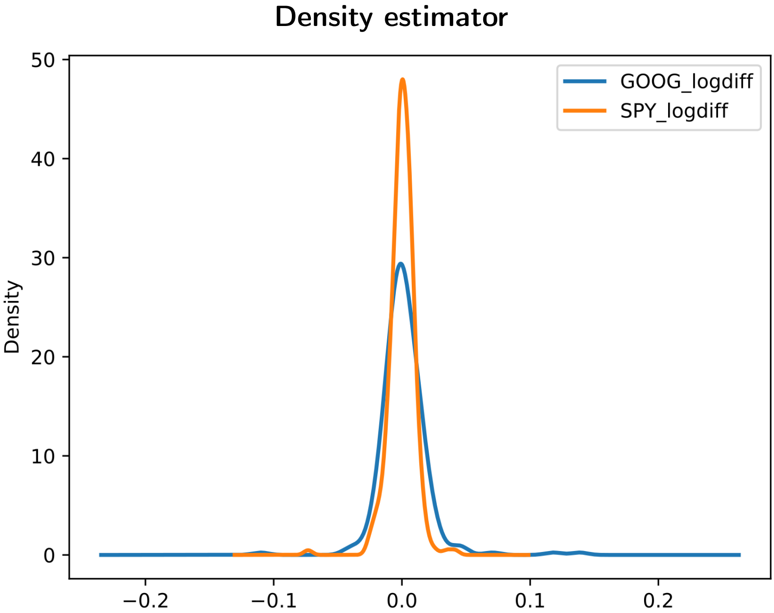
- Let G_t = price of GOOG, S_t = price of SPY.
- Define $\Delta_t \triangleq \log \frac{G_t}{G_{t-1}} - \log \frac{S_t}{S_{t-1}}$
- ... equals relative daily growth of Google stock w.r.t. S&P500
- Null hypothesis: $H : \mathbb{E}[\Delta] = 0$. Alternative $K : \mathbb{E}[\Delta] \neq 0$

Note: Before running t-test, check that Δ has nice distribution

Distribution of daily differences Δ_t



Distribution of daily differences Δ_t



Running t-test

Hypothesis testing setup

- Let G_t = price of GOOG, S_t = price of SPY.
- Define $\Delta_t \triangleq \log \frac{G_t}{G_{t-1}} - \log \frac{S_t}{S_{t-1}}$
- Null hypothesis: $H : \mathbb{E}[\Delta] = 0$. Alternative $K : \mathbb{E}[\Delta] \neq 0$

- Data: $n = 252$ daily values
- Empirical mean $\hat{\mathbb{E}}[\Delta_t] = 0.0014$ (or 0.14% relative daily growth)
- Empirical std. deviation: $\sqrt{\widehat{\sigma^2}} = 0.015$ (or $\pm 1.5\%$ r.d.g.)

$$\boxed{t_{obs} = \sqrt{\frac{n}{\widehat{\sigma^2}}}\hat{\mu} \approx 1.54}$$

Running t-test

Hypothesis testing setup

- Let G_t = price of GOOG, S_t = price of SPY.
- Define $\Delta_t \triangleq \log \frac{G_t}{G_{t-1}} - \log \frac{S_t}{S_{t-1}}$
- Null hypothesis: $H : \mathbb{E}[\Delta] = 0$. Alternative $K : \mathbb{E}[\Delta] \neq 0$

- Data: $n = 252$ daily values
- Empirical mean $\hat{\mathbb{E}}[\Delta_t] = 0.0014$ (or 0.14% relative daily growth)
- Empirical std. deviation: $\sqrt{\widehat{\sigma^2}} = 0.015$ (or $\pm 1.5\%$ r.d.g.)

$$\boxed{t_{obs} = \sqrt{\frac{n}{\widehat{\sigma^2}}}\hat{\mu} \approx 1.54}$$

- Under null $\mathbb{P}[T > t_{obs}] \approx 2\Phi(t_{obs}) \approx 0.12$
- ... about 12% chance to see this data under null.

Running t-test

Hypothesis testing setup

- Let G_t = price of GOOG, S_t = price of SPY.
- Define $\Delta_t \triangleq \log \frac{G_t}{G_{t-1}} - \log \frac{S_t}{S_{t-1}}$
- Null hypothesis: $H : \mathbb{E}[\Delta] = 0$. Alternative $K : \mathbb{E}[\Delta] \neq 0$

- Data: $n = 252$ daily values
- Empirical mean $\hat{\mathbb{E}}[\Delta_t] = 0.0014$ (or 0.14% relative daily growth)
- Empirical std. deviation: $\sqrt{\widehat{\sigma^2}} = 0.015$ (or $\pm 1.5\%$ r.d.g.)

$$\boxed{t_{obs} = \sqrt{\frac{n}{\widehat{\sigma^2}}}\hat{\mu} \approx 1.54}$$

- Under null $\mathbb{P}[T > t_{obs}] \approx 2\Phi(t_{obs}) \approx 0.12$
- ... If GOOG were SPY with larger σ^2, 12% chance to see this return

Running t-test

Hypothesis testing setup

- Let G_t = price of GOOG, S_t = price of SPY.
- Define $\Delta_t \triangleq \log \frac{G_t}{G_{t-1}} - \log \frac{S_t}{S_{t-1}}$
- Null hypothesis: $H : \mathbb{E}[\Delta] = 0$. Alternative $K : \mathbb{E}[\Delta] \neq 0$

- Data: $n = 252$ daily values
- Empirical mean $\hat{\mathbb{E}}[\Delta_t] = 0.0014$ (or 0.14% relative daily growth)
- Empirical std. deviation: $\sqrt{\widehat{\sigma^2}} = 0.015$ (or $\pm 1.5\%$ r.d.g.)

$$\boxed{t_{obs} = \sqrt{\frac{n}{\widehat{\sigma^2}}}\hat{\mu} \approx 1.54}$$

- Under null $\mathbb{P}[T > t_{obs}] \approx 2\Phi(t_{obs}) \approx 0.12$
- ... If GOOG were SPY with larger σ^2, 12% chance to see this return
- **NOT SIGNIFICANT**

Running t-test

Hypothesis testing setup

- Let G_t = price of GOOG, S_t = price of SPY.
- Define $\Delta_t \triangleq \log \frac{G_t}{G_{t-1}} - \log \frac{S_t}{S_{t-1}}$
- Null hypothesis: $H : \mathbb{E}[\Delta] = 0$. Alternative $K : \mathbb{E}[\Delta] \neq 0$

- Data: $n = 252$ daily values
- Empirical mean $\hat{\mathbb{E}}[\Delta_t] = 0.0014$ (or 0.14% relative daily growth)
- Empirical std. deviation: $\sqrt{\widehat{\sigma^2}} = 0.015$ (or $\pm 1.5\%$ r.d.g.)

$$\boxed{t_{obs} = \sqrt{\frac{n}{\widehat{\sigma^2}}}\hat{\mu} \approx 1.54}$$

- Under null $\mathbb{P}[T > t_{obs}] \approx 2\Phi(t_{obs}) \approx 0.12$
- ... If GOOG were SPY with larger σ^2, 12% chance to see this return
- **NOT SIGNIFICANT**
- AMZN grew by 120%. Significant: only 1% chance under null.

$\boxed{\text{The Wald test}}$

Wald test for parametric models

The next set of HT is:

- Have a good parametric model $X \sim P_\theta$ (e.g. Gaussian, $\theta = (\mu, \sigma)$)
- Want to test:

$$H : \theta \in \Theta_0 \qquad \text{vs.} \qquad K : \theta \notin \Theta_0$$

- Idea: Use (some) $\hat{\theta}$ and just check $\hat{\theta} \in \Theta_0$

Wald test for parametric models

The next set of HT is:

- Have a good parametric model $X \sim P_\theta$ (e.g. Gaussian, $\theta = (\mu, \sigma)$)
- Want to test:

$$H : \theta \in \Theta_0 \qquad \text{vs.} \qquad K : \theta \notin \Theta_0$$

- Idea: Use (some) $\hat{\theta}$ and just check $\hat{\theta} \in \Theta_0$
- Problems: What if Θ_0 is a point? How to adjust significance?

Wald test for parametric models

The next set of HT is:

- Have a good parametric model $X \sim P_\theta$ (e.g. Gaussian, $\theta = (\mu, \sigma)$)
- Want to test:

$$H : \theta \in \Theta_0 \qquad \text{vs.} \qquad K : \theta \notin \Theta_0$$

- Idea: Use (some) $\hat{\theta}$ and just check $\hat{\theta} \in \Theta_0$
- Problems: What if Θ_0 is a point? How to adjust significance?
- **Special case:** θ – scalar, $\Theta_0 = \{\theta_0\}$:

$$H : \theta = \theta_0 \qquad \text{vs.} \qquad K : \theta \neq \theta_0 \, .$$

The Wald test-statistic

$$W = \frac{\hat{\theta} - \theta_0}{\widehat{se}}$$

$\hat{\theta} =$ parameter estimator (eg. MLE),

$\widehat{se} =$ estimate of $\sqrt{\mathbb{E}^{\theta_0}[(\hat{\theta} - \theta_0)^2]}$

Wald test for parametric models

The next set of HT is:

- Have a good parametric model $X \sim P_\theta$ (e.g. $X \sim \mathrm{Ber}(\theta)$)
- **Special case:** θ – scalar, $\Theta_0 = \{\theta_0\}$:

$$H : \theta = \theta_0 \qquad \text{vs.} \qquad K : \theta \neq \theta_0 \,.$$

The Wald test-statistic

$$W = \frac{\hat{\theta} - \theta_0}{\widehat{se}}$$

$\hat{\theta}$ = parameter estimator (eg. MLE),

$\widehat{se}$ = estimate of std.err. (e.g. $\widehat{se} = \sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}}$)

- Assuming a) asymptotic normality of $\hat{\theta}$, b) consistency of $\hat{V}$:

$$W \approx \mathcal{N}(0,1) \qquad \text{for large } n$$

- So the Wald test (two-sided): If $|W| > z_{\frac{\alpha}{2}}$ then REJECT.

Wald test for parametric models

The next set of HT is:

- Have a good parametric model $X \sim P_\theta$ (e.g. $X \sim \mathrm{Ber}(\theta)$)
- **Special case:** θ – scalar, $\Theta_0 = \{\theta_0\}$:

$$H : \theta = \theta_0 \qquad \text{vs.} \qquad K : \theta \neq \theta_0 \,.$$

The Wald test-statistic

$$W = \frac{\hat{\theta} - \theta_0}{\widehat{se}}$$

$\hat{\theta}$ = parameter estimator (eg. MLE),

$\widehat{se}$ = estimate of std.err. (e.g. $\widehat{se} = \sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}}$)

- Assuming a) asymptotic normality of $\hat{\theta}$, b) consistency of $\hat{V}$:

$$W \approx \mathcal{N}(0,1) \qquad \text{for large } n$$

- So the Wald test (two-sided): If $|W| > z_{\frac{\alpha}{2}}$ then REJECT.
- General Wald: If $\theta_0 \notin \mathtt{CI}_{1-\alpha}$ then REJECT

Summary

Hypothesis testing mindset:

- Formulate hypotheses as **two collections of distributions**
- Come up with statistic whose distribution under H is known (or approximately known)
- Threshold statistic s.t. $\mathbb{P}[\text{reject}|H] \leq \alpha$ for pre-specified α.
- Only then see the data

Summary

Hypothesis testing mindset:

- Formulate hypotheses as **two collections of distributions**
- Come up with statistic whose distribution under H is known (or approximately known)
- Threshold statistic s.t. $\mathbb{P}[\text{reject}|H] \leq \alpha$ for pre-specified α.
- Only then see the data

Key lessons today:

- For null $\mathbb{E}[X] = \mu_0$: Compare empirical mean $\hat{\mu}$ to μ_0
- Reject null if $|\hat{\mu} - \mu_0|$ larger than $1.96 \cdot \mathrm{std}(\hat{\mu})$
- If $\mathrm{std}(\hat{\mu})$ is unknown, use $\widehat{se}(\hat{\mu})$
- ... trick known as Studentization

Summary

Hypothesis testing mindset:

- Formulate hypotheses as **two collections of distributions**
- Come up with statistic whose distribution under H is known (or approximately known)
- Threshold statistic s.t. $\mathbb{P}[\text{reject}|H] \leq \alpha$ for pre-specified α.
- Only then see the data

AGAIN: hypothesis $=$ collection of distributions

Key less

- For
- Reject null if $|\hat{\mu} - \mu_0|$ larger than $1.96 \cdot \mathrm{std}(\hat{\mu})$
- If $\mathrm{std}(\hat{\mu})$ is unknown, use $\widehat{se}(\hat{\mu})$
- ... trick known as Studentization