Statistics for Astronomers: Lecture 12, 2020.11.23

Prof. Sundar Srinivasan

IRyA/UNAM

Review

Empirical distribution function.

Bootstrap.

For bootstrap CIs, good discussion here.

Hypothesis testing

References

Barlow
AstroML

Lady Tasting Tea

Ronald Fisher offered [Blanche Muriel] Bristol a cup of hot tea that he had just drawn from an urn. Bristol declined it, saying that she preferred the flavour when the milk was poured into the cup before the tea. Fisher scoffed that the order of pouring could not affect the flavour. Bristol insisted that it did and that she could tell the difference. Overhearing this debate, William Roach said, 'Let's test her'.

– "Lady Tasting Tea", Rod Sturdivant.

Lady Tasting Tea

Ronald Fisher offered [Blanche Muriel] Bristol a cup of hot tea that he had just drawn from an urn. Bristol declined it, saying that she preferred the flavour when the milk was poured into the cup before the tea. Fisher scoffed that the order of pouring could not affect the flavour. Bristol insisted that it did and that she could tell the difference. Overhearing this debate, William Roach said, 'Let's test her'.

– "Lady Tasting Tea", Rod Sturdivant.

Experimental setup: prepare 8 cups of tea, 4 of which have milk poured in before tea.

Lady Tasting Tea

Ronald Fisher offered [Blanche Muriel] Bristol a cup of hot tea that he had just drawn from an urn. Bristol declined it, saying that she preferred the flavour when the milk was poured into the cup before the tea. Fisher scoffed that the order of pouring could not affect the flavour. Bristol insisted that it did and that she could tell the difference. Overhearing this debate, William Roach said, 'Let's test her'.

– "Lady Tasting Tea", Rod Sturdivant.

Experimental setup: prepare 8 cups of tea, 4 of which have milk poured in before tea.

Null hypothesis: Subject has no special ability. Test statistic: # cups successfully characterised.

Lady Tasting Tea

Ronald Fisher offered [Blanche Muriel] Bristol a cup of hot tea that he had just drawn from an urn. Bristol declined it, saying that she preferred the flavour when the milk was poured into the cup before the tea. Fisher scoffed that the order of pouring could not affect the flavour. Bristol insisted that it did and that she could tell the difference. Overhearing this debate, William Roach said, 'Let's test her'.

– "Lady Tasting Tea", Rod Sturdivant.

Experimental setup: prepare 8 cups of tea, 4 of which have milk poured in before tea.
Null hypothesis: Subject has no special ability. Test statistic: # cups successfully characterised.

$P(\geq 3 \text{ of } 4 \text{ cups correct by chance})$: $(16 + 1)/70 \approx 23\%$.
$P(4 \text{ of } 4 \text{ cups correct by chance})$: $1/70 \approx 1.4\% < 5\%$.

Lady Tasting Tea

Ronald Fisher offered [Blanche Muriel] Bristol a cup of hot tea that he had just drawn from an urn. Bristol declined it, saying that she preferred the flavour when the milk was poured into the cup before the tea. Fisher scoffed that the order of pouring could not affect the flavour. Bristol insisted that it did and that she could tell the difference. Overhearing this debate, William Roach said, 'Let's test her'.

– "Lady Tasting Tea", Rod Sturdivant.

Experimental setup: prepare 8 cups of tea, 4 of which have milk poured in before tea.

Null hypothesis: Subject has no special ability. Test statistic: # cups successfully characterised.

$P(\geq 3 \text{ of } 4 \text{ cups correct by chance})$: $(16 + 1)/70 \approx 23\%$.

$P(4 \text{ of } 4 \text{ cups correct by chance})$: $1/70 \approx 1.4\% < 5\%$.

Critical region for rejection of null hypothesis: 4 out of 4 possible cups successfully characterised.

$P(4 \text{ of } 4 \text{ cups correctly characterised})$ $(1/70 \approx 1.4\% < 5\%)$.

Lady Tasting Tea

Ronald Fisher offered [Blanche Muriel] Bristol a cup of hot tea that he had just drawn from an urn. Bristol declined it, saying that she preferred the flavour when the milk was poured into the cup before the tea. Fisher scoffed that the order of pouring could not affect the flavour. Bristol insisted that it did and that she could tell the difference. Overhearing this debate, William Roach said, 'Let's test her'.

– "Lady Tasting Tea", Rod Sturdivant.

Experimental setup: prepare 8 cups of tea, 4 of which have milk poured in before tea.

Null hypothesis: Subject has no special ability. Test statistic: # cups successfully characterised.

$P(\geq 3 \text{ of } 4 \text{ cups correct by chance})$: $(16 + 1)/70 \approx 23\%$.

$P(4 \text{ of } 4 \text{ cups correct by chance})$: $1/70 \approx 1.4\% < 5\%$.

Critical region for rejection of null hypothesis: 4 out of 4 possible cups successfully characterised.

$P(4 \text{ of } 4 \text{ cups correctly characterised})$ $(1/70 \approx 1.4\% < 5\%)$.

Bristol correctly characterised all eight cups.

Introduction

"Do the data provide sufficient evidence to conclude that we must depart from our original assumption concerning the state of nature?"

– J. C. Watkins, *An Introduction to the Science of Statistics*.

Introduction

"Do the data provide sufficient evidence to conclude that we must depart from our original assumption concerning the state of nature?"

– J. C. Watkins, *An Introduction to the Science of Statistics*.

What kinds of questions can be answered? (from Barlow)

~~What is the straight line fit for y vs. x?~~ *Does* y increase with x?

~~What is the strength of the effect?~~ *Is* the effect present?

~~What are the values of a and b?~~ *Do* a and b have the same value?

Formulate the question precisely by expressing it as a hypothesis.

Introduction

"Do the data provide sufficient evidence to conclude that we must depart from our original assumption concerning the state of nature?"

– J. C. Watkins, *An Introduction to the Science of Statistics*.

What kinds of questions can be answered? (from Barlow)

- ~~What is the straight line fit for y vs. x?~~ *Does* y increase with x?
- ~~What is the strength of the effect?~~ *Is* the effect present?
- ~~What are the values of a and b?~~ *Do* a and b have the same value?

Formulate the question precisely by expressing it as a hypothesis.

Statistical test: Procedure. Input: samples. Computes: test statistic. Output: a hypothesis.

Hypothesis: assertion/statement that can be tested using observations (*e.g.*, "the population mean is < 5").

Introduction

"Do the data provide sufficient evidence to conclude that we must depart from our original assumption concerning the state of nature?"

– J. C. Watkins, *An Introduction to the Science of Statistics*.

What kinds of questions can be answered? (from Barlow)

- ~~What is the straight line fit for y vs. x?~~ *Does* y increase with x?
- ~~What is the strength of the effect?~~ *Is* the effect present?
- ~~What are the values of a and b?~~ *Do* a and b have the same value?

Formulate the question precisely by expressing it as a hypothesis.

Statistical test: Procedure. Input: samples. Computes: test statistic. Output: a hypothesis.

Hypothesis: assertion/statement that can be tested using observations (*e.g.*, "the population mean is < 5").

- Can be simple (complete description of the underlying population distribution
  - *e.g.*, "the errors are Gaussian with mean 0 and variance 1")

Introduction

"Do the data provide sufficient evidence to conclude that we must depart from our original assumption concerning the state of nature?"

– J. C. Watkins, *An Introduction to the Science of Statistics*.

What kinds of questions can be answered? (from Barlow)

- ~~What is the straight line fit for y vs. x?~~ *Does* y increase with x?
- ~~What is the strength of the effect?~~ *Is* the effect present?
- ~~What are the values of a and b?~~ *Do* a and b have the same value?

Formulate the question precisely by expressing it as a hypothesis.

Statistical test: Procedure. Input: samples. Computes: test statistic. Output: a hypothesis.

Hypothesis: assertion/statement that can be tested using observations (*e.g.*, "the population mean is < 5").

- Can be simple (complete description of the underlying population distribution
  - *e.g.*, "the errors are Gaussian with mean 0 and variance 1")
- or composite (underlying population distribution unclear)
  - *e.g.*, "the mean is not 0".

Introduction

"Do the data provide sufficient evidence to conclude that we must depart from our original assumption concerning the state of nature?"

– J. C. Watkins, *An Introduction to the Science of Statistics*.

What kinds of questions can be answered? (from Barlow)

~~What is the straight line fit for y vs. x?~~	*Does* y increase with x?
~~What is the strength of the effect?~~	*Is* the effect present?
~~What are the values of a and b?~~	*Do* a and b have the same value?

Formulate the question precisely by expressing it as a hypothesis.

Statistical test: Procedure. Input: samples. Computes: test statistic. Output: a hypothesis.

Hypothesis: assertion/statement that can be tested using observations (*e.g.*, "the population mean is < 5").

- Can be simple (complete description of the underlying population distribution
  - *e.g.*, "the errors are Gaussian with mean 0 and variance 1")
- or composite (underlying population distribution unclear)
  - *e.g.*, "the mean is not 0".
- Can be two-tailed/non-directional test *e.g.*, "$\theta = \theta_0$", "$-5 \leq \mu \leq 5$".
- or one-tailed/directional test *e.g.*, "$\theta > \theta_0$", "$\mu < 5$".

The Null Hypothesis H_0

Typically, a statement expressing lack of correlation between observations and the suggested model (*i.e.*, the data are not significantly different from noise), and the alternate hypothesis H_A suggests a relationship.

Want to demonstrate that <effect> exists? Start by stating it doesn't, then find out whether data provides enough evidence to reject H_0 – hypothesis testing.

The Null Hypothesis H_0

Typically, a statement expressing lack of correlation between observations and the suggested model (*i.e.*, the data are not significantly different from noise), and the alternate hypothesis H_A suggests a relationship.

Want to demonstrate that <effect> exists? Start by stating it doesn't, then find out whether data provides enough evidence to reject H_0 – hypothesis testing.

"*[The Null Hypothesis is] never proved or established, but is possibly disproved, in the course of experimentation.*"
– R. A. Fisher.

The Null Hypothesis H_0

Typically, a statement expressing lack of correlation between observations and the suggested model (*i.e.*, the data are not significantly different from noise), and the alternate hypothesis H_A suggests a relationship.

Want to demonstrate that <effect> exists? Start by stating it doesn't, then find out whether data provides enough evidence to reject H_0 – hypothesis testing.

"*[The Null Hypothesis is] never proved or established, but is possibly disproved, in the course of experimentation.*"
– R. A. Fisher.

> Also, 28 days after they started treatment, 10.4% of those treated with hydroxychloroquine died, just slightly lower than the 10.6% fatality rate in the placebo group.
>
> "The results show that hydroxychloroquine did not help patients recover from COVID-19," study co-author Dr. Wesley H. Self told UPI.

W. H. Self, et al. *JAMA*, 10.1001/jama.2020.22240.

The Null Hypothesis H_0

Typically, a statement expressing lack of correlation between observations and the suggested model (*i.e.*, the data are not significantly different from noise), and the alternate hypothesis H_A suggests a relationship.

Want to demonstrate that <effect> exists? Start by stating it doesn't, then find out whether data provides enough evidence to reject H_0 – hypothesis testing.

"*[The Null Hypothesis is] never proved or established, but is possibly disproved, in the course of experimentation.*"
– R. A. Fisher.

Also, 28 days after they started treatment, 10.4% of those treated with hydroxychloroquine died, just slightly lower than the 10.6% fatality rate in the placebo group.

"The results show that hydroxychloroquine did not help patients recover from COVID-19," study co-author Dr. Wesley H. Self told UPI.

W. H. Self, et al. *JAMA*, 10.1001/jama.2020.22240.

If the probability of the data occurring by chance is below a threshold (significance), then we reject the null hypothesis.

The Null Hypothesis H_0

Typically, a statement expressing lack of correlation between observations and the suggested model (*i.e.*, the data are not significantly different from noise), and the alternate hypothesis H_A suggests a relationship.

Want to demonstrate that <effect> exists? Start by stating it doesn't, then find out whether data provides enough evidence to reject H_0 – hypothesis testing.

"*[The Null Hypothesis is] never proved or established, but is possibly disproved, in the course of experimentation.*"
– R. A. Fisher.

> Also, 28 days after they started treatment, 10.4% of those treated with hydroxychloroquine died, just slightly lower than the 10.6% fatality rate in the placebo group.
>
> "The results show that hydroxychloroquine did not help patients recover from COVID-19," study co-author Dr. Wesley H. Self told UPI.

W. H. Self, et al. *JAMA*, 10.1001/jama.2020.22240.

If the probability of the data occurring by chance is below a threshold (significance), then we reject the null hypothesis.

Frequentist inference: probability that a given hypothesis is correct is either 0 or 1.
Just because we reject H_0 on the basis of one dataset doesn't mean H_0 is wrong or H_A is correct.
At 95% confidence, frequentist procedure will reject H_0 for 5% of datasets drawn from H_0!

Type I and II Errors

Choose null (H_0) and alternate (H_A or H_1) hypotheses.

Compute significance (α) using data.

- α is the level of tolerance for incorrectly rejecting H_0.
- Outcome "significant" if small chance of occurrence from H_0.

Given α, only two possible outcomes: reject/unable to reject H_0.

Type I and II Errors

Choose null (H_0) and alternate (H_A or H_1) hypotheses.

Compute significance (α) using data.

- α is the level of tolerance for incorrectly rejecting H_0.
- Outcome "significant" if small chance of occurrence from H_0.

Given α, only two possible outcomes: reject/unable to reject H_0.

Table of error types		Null hypothesis (H_0) is	
		True	False
Decision about null hypothesis (H_0)	Don't reject	Correct inference (true negative) (probability = $1-\alpha$)	Type II error (false negative) (probability = β)
	Reject	Type I error (false positive) (probability = α)	Correct inference (true positive) (probability = $1-\beta$)

Source: Wikipedia

Type I and II Errors

Choose null (H_0) and alternate (H_A or H_1) hypotheses.

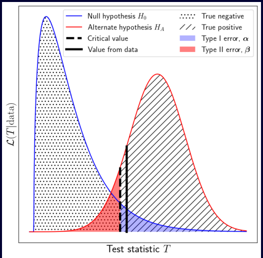
Compute significance (α) using data.

- α is the level of tolerance for incorrectly rejecting H_0.
- Outcome "significant" if small chance of occurrence from H_0.

Given α, only two possible outcomes: reject/unable to reject H_0.

Table of error types		Null hypothesis (H_0) is	
		True	False
Decision about null hypothesis (H_0)	Don't reject	Correct inference (true negative) (probability = $1-\alpha$)	Type II error (false negative) (probability = β)
	Reject	Type I error (false positive) (probability = α)	Correct inference (true positive) (probability = $1-\beta$)

Source: Wikipedia



"H_0 rejected at level α for these data."

Type I and II Errors

Choose null (H_0) and alternate (H_A or H_1) hypotheses.

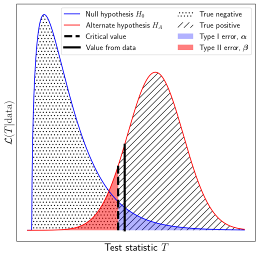
Compute significance (α) using data.

- α is the level of tolerance for incorrectly rejecting H_0.
- Outcome "significant" if small chance of occurrence from H_0.

Given α, only two possible outcomes: reject/unable to reject H_0.

Table of error types		Null hypothesis (H_0) is True	Null hypothesis (H_0) is False
Decision about null hypothesis (H_0)	Don't reject	Correct inference (true negative) (probability = 1−α)	Type II error (false negative) (probability = β)
	Reject	Type I error (false positive) (probability = α)	Correct inference (true positive) (probability = 1−β)

Source: Wikipedia



"H_0 rejected at level α for these data."

Acceptance region: set of test statistic values for which we fail to reject H_0.

Type I and II Errors

Choose null (H_0) and alternate (H_A or H_1) hypotheses.

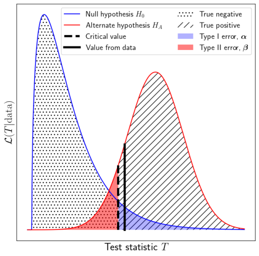
Compute significance (α) using data.

- α is the level of tolerance for incorrectly rejecting H_0.
- Outcome "significant" if small chance of occurrence from H_0.

Given α, only two possible outcomes: reject/unable to reject H_0.

Table of error types		Null hypothesis (H_0) is True	Null hypothesis (H_0) is False
Decision about null hypothesis (H_0)	Don't reject	Correct inference (true negative) (probability = 1−α)	Type II error (false negative) (probability = β)
	Reject	Type I error (false positive) (probability = α)	Correct inference (true positive) (probability = 1−β)

Source: Wikipedia



"H_0 rejected at level α for these data."

Acceptance region: set of test statistic values for which we fail to reject H_0.

Rejection or critical region: set of test statistic values for which we are able to reject H_0.

Type I and II Errors

Choose null (H_0) and alternate (H_A or H_1) hypotheses.

Compute significance (α) using data.

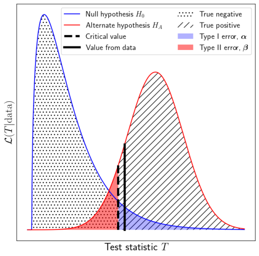
α is the level of tolerance for incorrectly rejecting H_0.

Outcome "significant" if small chance of occurrence from H_0.

Given α, only two possible outcomes: reject/unable to reject H_0.

Table of error types		Null hypothesis (H_0) is	
		True	False
Decision about null hypothesis (H_0)	Don't reject	Correct inference (true negative) (probability = $1-\alpha$)	Type II error (false negative) (probability = β)
	Reject	Type I error (false positive) (probability = α)	Correct inference (true positive) (probability = $1-\beta$)

Source: Wikipedia



"H_0 rejected at level α for these data."

Acceptance region: set of test statistic values for which we fail to reject H_0.

Rejection or critical region: set of test statistic values for which we are able to reject H_0.

Critical value: the threshold separating acceptance and rejection regions.

Type I and II Errors

Choose null (H_0) and alternate (H_A or H_1) hypotheses.

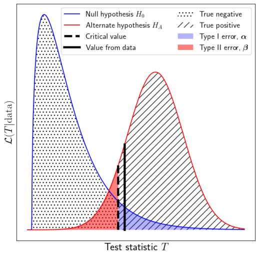
Compute significance (α) using data.

- α is the level of tolerance for incorrectly rejecting H_0.
- Outcome "significant" if small chance of occurrence from H_0.

Given α, only two possible outcomes: reject/unable to reject H_0.

Table of error types		Null hypothesis (H_0) is	
		True	False
Decision about null hypothesis (H_0)	Don't reject	Correct inference (true negative) (probability = 1−α)	Type II error (false negative) (probability = β)
	Reject	Type I error (false positive) (probability = α)	Correct inference (true positive) (probability = 1−β)

Source: Wikipedia



"H_0 rejected at level α for these data."

Acceptance region: set of test statistic values for which we fail to reject H_0.

Rejection or critical region: set of test statistic values for which we are able to reject H_0.

Critical value: the threshold separating acceptance and rejection regions.

p-value: Assuming H_0 is true, the probability of observing a result at least as extreme as the observed value of the test statistic.

Type I and II Errors

Choose null (H_0) and alternate (H_A or H_1) hypotheses.

Compute significance (α) using data.

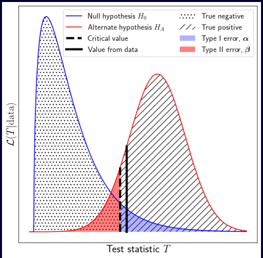
  α is the level of tolerance for incorrectly rejecting H_0.

  Outcome "significant" if small chance of occurrence from H_0.

Given α, only two possible outcomes: reject/unable to reject H_0.

Table of error types		Null hypothesis (H_0) is	
		True	False
Decision about null hypothesis (H_0)	Don't reject	Correct inference (true negative) (probability = 1−α)	Type II error (false negative) (probability = β)
	Reject	Type I error (false positive) (probability = α)	Correct inference (true positive) (probability = 1−β)

Source: Wikipedia



"H_0 rejected at level α for these data."

Acceptance region: set of test statistic values for which we fail to reject H_0.

Rejection or critical region: set of test statistic values for which we are able to reject H_0.

Critical value: the threshold separating acceptance and rejection regions.

p-value: Assuming H_0 is true, the probability of observing a result at least as extreme as the observed value of the test statistic.

Error rates:

  Type I (false +ve, false alarm): $P(\text{reject } H_0 \mid H_0 \text{ true}) \equiv \alpha$.

  Type II (false −ve): $P(\text{don't reject } H_0 \mid H_0 \text{ false}) \equiv \beta$.

Type I/II errors and classification algorithms

Classification typically involves placing boundaries in multidimensional parameter space to separate "clusters" of objects.

Type I/II errors and classification algorithms

Classification typically involves placing boundaries in multidimensional parameter space to separate "clusters" of objects.

Example: YSO researcher wants to identify promising massive embedded YSO candidates for spectroscopic follow-up. She devises cuts in colour-magnitude and colour-colour space to separate "high-reliability" YSO candidates from other kinds of sources with similar colours (*e.g.*, highly evolved dusty AGB stars, background galaxies).

Type I/II errors and classification algorithms

Classification typically involves placing boundaries in multidimensional parameter space to separate "clusters" of objects.

Example: YSO researcher wants to identify promising massive embedded YSO candidates for spectroscopic follow-up. She devises cuts in colour-magnitude and colour-colour space to separate "high-reliability" YSO candidates from other kinds of sources with similar colours (*e.g.*, highly evolved dusty AGB stars, background galaxies).

Type I error = false +ve = contamination ("spurious detections") to the YSO candidate sample.
Type II error = false −ve rate ("missed sources") reduces the completeness of the YSO candidate sample.

Type I/II errors and classification algorithms

Classification typically involves placing boundaries in multidimensional parameter space to separate "clusters" of objects.

Example: YSO researcher wants to identify promising massive embedded YSO candidates for spectroscopic follow-up. She devises cuts in colour-magnitude and colour-colour space to separate "high-reliability" YSO candidates from other kinds of sources with similar colours (*e.g.*, highly evolved dusty AGB stars, background galaxies).

Type I error = false +ve = contamination ("spurious detections") to the YSO candidate sample.
Type II error = false −ve rate ("missed sources") reduces the completeness of the YSO candidate sample.
Compromise between increasing completeness and decreasing contamination – received operating characteristic (ROC) curve (true +ve rate vs. true −ve rate).

See Sec. 4.6.1 in the AstroML book.

Hypothesis testing: basic procedure

1. Identify a null hypothesis and an alternate hypothesis, choose significance threshold α.
2. Design test statistic T. Assuming H_0 is true, obtain the distribution of T.
   Usually complicated/unknown; use the asymptotic distribution ($N \to \infty$).
3. Using the data, compute t, the observed value of T.
4. Compute the p-value: $p \equiv P(T = t | H_0 \text{ is true})$.
5. If the $p < \alpha$, the tolerance for false negatives, reject H_0 at significance level α.

Hypothesis testing: basic procedure

1. Identify a null hypothesis and an alternate hypothesis, choose significance threshold α.
2. Design test statistic T. Assuming H_0 is true, obtain the distribution of T.
   Usually complicated/unknown; use the asymptotic distribution ($N \to \infty$).
3. Using the data, compute t, the observed value of T.
4. Compute the p-value: $p \equiv P(T = t | H_0 \text{ is true})$.
5. If the $p < \alpha$, the tolerance for false negatives, reject H_0 at significance level α.

Example 1

Observation: Tossing a coin 10 times, we observe 9 heads.

Statistic: S_{10}, the total number of heads in 10 tosses.

H_0: fair coin. Under H_0, $S_{10} \sim \text{Binomial}(1/2)$.

H_A: $p \neq 0.5$ (two-tailed).

Significance chosen: $\alpha = 0.05$.

Hypothesis testing: basic procedure

1. Identify a null hypothesis and an alternate hypothesis, choose significance threshold α.
2. Design test statistic T. Assuming H_0 is true, obtain the distribution of T.
   Usually complicated/unknown; use the asymptotic distribution ($N \to \infty$).
3. Using the data, compute t, the observed value of T.
4. Compute the p-value: $p \equiv P(T = t | H_0 \text{ is true})$.
5. If the $p < \alpha$, the tolerance for false negatives, reject H_0 at significance level α.

Example 1

Observation: Tossing a coin 10 times, we observe 9 heads.

Statistic: S_{10}, the total number of heads in 10 tosses.

H_0: fair coin. Under H_0, $S_{10} \sim \text{Binomial}(1/2)$.

H_A: $p \neq 0.5$ (two-tailed).

Significance chosen: $\alpha = 0.05$.

p-value: $P(S_{10} \geq 9) = \binom{10}{9}\frac{1}{2}^{10} + \binom{10}{10}\frac{1}{2}^{10} \approx 0.0098$.

Since the p-value $(= 0.009) <$ significance, reject H_0 at significance level $\alpha = 0.05$.

Hypothesis testing contd.

Example 2 (Barlow 8.2.2)
55% of patients suffering from a disease are spontaneously cured within a week.
A new medication is tested on 105 patients. How many patients need to be cured in a week to decide whether the medication is effective at 5% significance?

H_0: $p \leq 0.55$; H_A: $p > 0.55$ (one-tailed test)

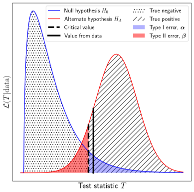
Statistic: k, the total number of people cured within a week.
$k \sim \mathrm{Binomial}(0.55)$ under null hypothesis.

Significance chosen: $\alpha = 0.05$.

We are looking for k_α such that $P(k \geq k_\alpha) < \alpha$.

"To reject H_0 at 5% significance, more than ( ) patients need to be cured within a week."

Statistical Power and the likelihood-ratio test

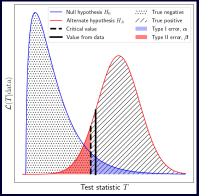


Statistical power of a test: $P(\text{reject } H_0 \mid H_0 \text{ false}) \equiv 1 - \beta$

As critical/threshold value $\uparrow$, $\alpha \downarrow$ but power also $\downarrow$.

Efficiency of a test: sample size required to achieve a given power.

Statistical Power and the likelihood-ratio test



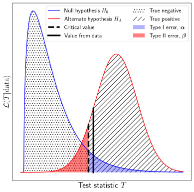
Statistical power of a test: $P(\text{reject } H_0 \mid H_0 \text{ false}) \equiv 1 - \beta$

As critical/threshold value ↑, α ↓ but power also ↓.

Efficiency of a test: sample size required to achieve a given power.

Ideal situation: maximum power for a given α. Not possible in general. (*e.g.*, unknown or complicated distribution, composite hypotheses).

Statistical Power and the likelihood-ratio test



Statistical power of a test: $P(\text{reject } H_0 \mid H_0 \text{ false}) \equiv 1 - \beta$

As critical/threshold value $\uparrow$, $\alpha \downarrow$ but power also $\downarrow$.

Efficiency of a test: sample size required to achieve a given power.

Ideal situation: maximum power for a given α. Not possible in general.

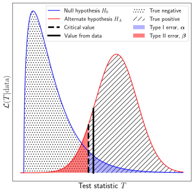
(*e.g.*, unknown or complicated distribution, composite hypotheses).

Neyman-Pearson Lemma

If both H_0 and H_A are simple, $p_T(t \mid H_0 \text{ true})$ and $p_T(t \mid H_A \text{ true})$ known.

$\Longrightarrow$ the likelihood ratio is the most powerful test statistic.

Statistical Power and the likelihood-ratio test



Statistical power of a test: $P(\text{reject } H_0 \mid H_0 \text{ false}) \equiv 1-\beta$

As critical/threshold value $\uparrow$, $\alpha \downarrow$ but power also $\downarrow$.

Efficiency of a test: sample size required to achieve a given power.

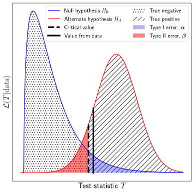
Ideal situation: maximum power for a given α. Not possible in general.
(*e.g.*, unknown or complicated distribution, composite hypotheses).

Neyman-Pearson Lemma

If both H_0 and H_A are simple, $p_T(t \mid H_0 \text{ true})$ and $p_T(t \mid H_A \text{ true})$ known.
$\Longrightarrow$ the likelihood ratio is the most powerful test statistic.

$$\text{Likelihood ratio} = \frac{\text{likelihood } H_A \text{ true given data}}{\text{likelihood } H_0 \text{ true given data}} > \text{threshold} \Longrightarrow \text{reject } H_0.$$

Statistical Power and the likelihood-ratio test



Statistical power of a test: $P(\text{reject } H_0 \mid H_0 \text{ false}) \equiv 1 - \beta$

As critical/threshold value $\uparrow$, $\alpha \downarrow$ but power also $\downarrow$.

Efficiency of a test: sample size required to achieve a given power.

Ideal situation: maximum power for a given α. Not possible in general.

(*e.g.*, unknown or complicated distribution, composite hypotheses).

Neyman-Pearson Lemma

If both H_0 and H_A are simple, $p_T(t \mid H_0 \text{ true})$ and $p_T(t \mid H_A \text{ true})$ known.

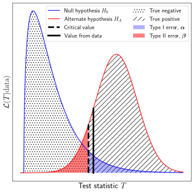
$\Longrightarrow$ the likelihood ratio is the most powerful test statistic.

$$\text{Likelihood ratio} = \frac{\text{likelihood } H_A \text{ true given data}}{\text{likelihood } H_0 \text{ true given data}} > \text{threshold} \Longrightarrow \text{reject } H_0.$$

If H_0, H_A simple, write in terms of parameter values: $LR = \dfrac{\mathscr{L}(\theta = \theta_1 \mid H_A \text{ true})}{\mathscr{L}(\theta = \theta_0 \mid H_0 \text{ true})} > \text{threshold}.$

The value of the threshold is picked such that the false-alarm probability is α.

Statistical Power and the likelihood-ratio test



Statistical power of a test: $P(\text{reject } H_0 \mid H_0 \text{ false}) \equiv 1 - \beta$

As critical/threshold value $\uparrow$, $\alpha \downarrow$ but power also $\downarrow$.

Efficiency of a test: sample size required to achieve a given power.

Ideal situation: maximum power for a given α. Not possible in general.
(*e.g.*, unknown or complicated distribution, composite hypotheses).

Neyman-Pearson Lemma

If both H_0 and H_A are simple, $p_T(t \mid H_0 \text{ true})$ and $p_T(t \mid H_A \text{ true})$ known.
$\implies$ the likelihood ratio is the most powerful test statistic.

$$\text{Likelihood ratio} = \frac{\text{likelihood } H_A \text{ true given data}}{\text{likelihood } H_0 \text{ true given data}} > \text{threshold} \implies \text{reject } H_0.$$

If H_0, H_A simple, write in terms of parameter values: $LR = \dfrac{\mathscr{L}(\theta = \theta_1 \mid H_A \text{ true})}{\mathscr{L}(\theta = \theta_0 \mid H_0 \text{ true})} > \text{threshold}.$

The value of the threshold is picked such that the false-alarm probability is α.

Typically, for convenience, written in terms of log-likelihood.

Recall: for Gaussian variables, $\ln \mathscr{L} = \text{constant} - \dfrac{1}{2}\chi^2$.

Wilks' Theorem: asymptotic behavior of $\ln LR$ under H_0 is χ^2!

Likelihood-ratio test example

$X_i(i = 1, \cdots, N) \sim \mathcal{N}(\mu, \sigma^2)$ with $\sigma = 4$ and μ unknown. H_0: $\mu = 0$, H_A: $\mu = 4$.

Find N and LR threshold such that we are able to reject H_0 at significance $\alpha = 0.05$ and our test has power $1 - \beta = 0.95$.

Likelihood-ratio test example

$X_i (i = 1, \cdots, N) \sim \mathcal{N}(\mu, \sigma^2)$ with $\sigma = 4$ and μ unknown. H_0: $\mu = 0$, H_A: $\mu = 4$.

Find N and LR threshold such that we are able to reject H_0 at significance $\alpha = 0.05$ and our test has power $1 - \beta = 0.95$.

For both hypotheses, $\mathscr{L}(\mu) = \prod_{i=1}^{N} \left(\frac{1}{\sqrt{2\pi}\sigma} \right)^N \exp\left[-\frac{1}{2} \left(\frac{x_i - \mu}{\sigma} \right)^2 \right] \Longrightarrow \ln \mathscr{L}(\mu) = \text{const.} - \frac{1}{2\sigma^2} \sum_{i=1}^{N} (x_i - \mu)^2$

Likelihood-ratio test example

$X_i(i = 1, \cdots, N) \sim \mathcal{N}(\mu, \sigma^2)$ with $\sigma = 4$ and μ unknown. H_0: $\mu = 0$, H_A: $\mu = 4$.

Find N and LR threshold such that we are able to reject H_0 at significance $\alpha = 0.05$ and our test has power $1 - \beta = 0.95$.

For both hypotheses, $\mathscr{L}(\mu) = \prod_{i=1}^{N} \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^N \exp\left[-\frac{1}{2}\left(\frac{x_i - \mu}{\sigma}\right)^2\right] \implies \ln \mathscr{L}(\mu) = \text{const.} - \frac{1}{2\sigma^2}\sum_{i=1}^{N}(x_i - \mu)^2$

$$\implies \ln LR = -\frac{1}{2\sigma^2}\sum_{i=1}^{N}\left((x_i - \mu_2)^2 - (x_i - \mu_1)^2\right) = -\frac{1}{2\sigma^2}\sum_{i=1}^{N}\left(2x_i(\mu_1 - \mu_2) + \mu_2^2 - \mu_1^2\right)$$

Likelihood-ratio test example

$X_i(i = 1, \cdots, N) \sim \mathcal{N}(\mu, \sigma^2)$ with $\sigma = 4$ and μ unknown. H_0: $\mu = 0$, H_A: $\mu = 4$.

Find N and LR threshold such that we are able to reject H_0 at significance $\alpha = 0.05$ and our test has power $1 - \beta = 0.95$.

For both hypotheses, $\mathscr{L}(\mu) = \prod_{i=1}^{N} \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^N \exp\left[-\frac{1}{2}\left(\frac{x_i - \mu}{\sigma}\right)^2\right] \Longrightarrow \ln \mathscr{L}(\mu) = \mathrm{const.} - \frac{1}{2\sigma^2}\sum_{i=1}^{N}(x_i - \mu)^2$

$$\Longrightarrow \ln LR = -\frac{1}{2\sigma^2}\sum_{i=1}^{N}\left((x_i - \mu_2)^2 - (x_i - \mu_1)^2\right) = -\frac{1}{2\sigma^2}\sum_{i=1}^{N}\left(2x_i(\mu_1 - \mu_2) + \mu_2^2 - \mu_1^2\right)$$

Plugging in $\mu_1 = 0$, $\mu_2 = 4$, $\ln LR = -\frac{1}{2\sigma^2}\sum_{i=1}^{N}\left(-8x_i + 16\right) = \frac{N}{4}\left(\bar{x} - 2\right)$

Likelihood-ratio test example

$X_i(i = 1, \cdots, N) \sim \mathcal{N}(\mu, \sigma^2)$ with $\sigma = 4$ and μ unknown. H_0: $\mu = 0$, H_A: $\mu = 4$.

Find N and LR threshold such that we are able to reject H_0 at significance $\alpha = 0.05$ and our test has power $1 - \beta = 0.95$.

For both hypotheses, $\mathscr{L}(\mu) = \prod_{i=1}^{N} \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^N \exp\left[-\frac{1}{2}\left(\frac{x_i - \mu}{\sigma}\right)^2\right] \implies \ln \mathscr{L}(\mu) = \mathrm{const.} - \frac{1}{2\sigma^2}\sum_{i=1}^{N}(x_i - \mu)^2$

$$\implies \ln LR = -\frac{1}{2\sigma^2}\sum_{i=1}^{N}\Big((x_i - \mu_2)^2 - (x_i - \mu_1)^2\Big) = -\frac{1}{2\sigma^2}\sum_{i=1}^{N}\Big(2x_i(\mu_1 - \mu_2) + \mu_2^2 - \mu_1^2\Big)$$

Plugging in $\mu_1 = 0, \mu_2 = 4$, $\ln LR = -\frac{1}{2\sigma^2}\sum_{i=1}^{N}\Big(-8x_i + 16\Big) = \frac{N}{4}\Big(\overline{x} - 2\Big)$

In order to reject H_0, we need $\ln LR = \frac{N}{4}\Big(\overline{x} - 2\Big) >$ some threshold. Since N is (an unknown) constant, we need $\overline{x} >$ some threshold c (say).

this makes sense – in order to distinguish the data from noise, its mean has to be > 0.

Likelihood-ratio test example

$X_i (i = 1, \cdots, N) \sim \mathcal{N}(\mu, \sigma^2)$ with $\sigma = 4$ and μ unknown. H_0: $\mu = 0$, H_A: $\mu = 4$.

Find N and LR threshold such that we are able to reject H_0 at significance $\alpha = 0.05$ and our test has power $1 - \beta = 0.95$.

For both hypotheses, $\mathscr{L}(\mu) = \prod_{i=1}^{N} \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^N \exp\left[-\frac{1}{2}\left(\frac{x_i - \mu}{\sigma}\right)^2\right] \implies \ln \mathscr{L}(\mu) = \mathrm{const.} - \frac{1}{2\sigma^2}\sum_{i=1}^{N}(x_i - \mu)^2$

$$\implies \ln LR = -\frac{1}{2\sigma^2}\sum_{i=1}^{N}\Big((x_i - \mu_2)^2 - (x_i - \mu_1)^2\Big) = -\frac{1}{2\sigma^2}\sum_{i=1}^{N}\Big(2x_i(\mu_1 - \mu_2) + \mu_2^2 - \mu_1^2\Big)$$

Plugging in $\mu_1 = 0, \mu_2 = 4$, $\ln LR = -\frac{1}{2\sigma^2}\sum_{i=1}^{N}\Big(-8x_i + 16\Big) = \frac{N}{4}\Big(\overline{x} - 2\Big)$

In order to reject H_0, we need $\ln LR = \frac{N}{4}\Big(\overline{x} - 2\Big) >$ some threshold. Since N is (an unknown) constant, we need $\overline{x} >$ some threshold c (say).

this makes sense – in order to distinguish the data from noise, its mean has to be > 0.

Recall: CLT means that $\overline{x} \sim \mathcal{N}(\mu, \sigma^2/N)$.

Likelihood-ratio test example

$X_i (i = 1, \cdots, N) \sim \mathcal{N}(\mu, \sigma^2)$ with $\sigma = 4$ and μ unknown. H_0: $\mu = 0$, H_A: $\mu = 4$.

Find N and LR threshold such that we are able to reject H_0 at significance $\alpha = 0.05$ and our test has power $1 - \beta = 0.95$.

For both hypotheses, $\mathscr{L}(\mu) = \prod_{i=1}^{N} \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^N \exp\left[-\frac{1}{2}\left(\frac{x_i - \mu}{\sigma}\right)^2\right] \Longrightarrow \ln \mathscr{L}(\mu) = \text{const.} - \frac{1}{2\sigma^2}\sum_{i=1}^{N}(x_i - \mu)^2$

$$\Longrightarrow \ln LR = -\frac{1}{2\sigma^2}\sum_{i=1}^{N}\left((x_i - \mu_2)^2 - (x_i - \mu_1)^2\right) = -\frac{1}{2\sigma^2}\sum_{i=1}^{N}\left(2x_i(\mu_1 - \mu_2) + \mu_2^2 - \mu_1^2\right)$$

Plugging in $\mu_1 = 0$, $\mu_2 = 4$, $\ln LR = -\frac{1}{2\sigma^2}\sum_{i=1}^{N}\left(-8x_i + 16\right) = \frac{N}{4}\left(\overline{x} - 2\right)$

In order to reject H_0, we need $\ln LR = \frac{N}{4}\left(\overline{x} - 2\right) >$ some threshold. Since N is (an unknown) constant, we need $\overline{x} >$ some threshold c (say).

this makes sense – in order to distinguish the data from noise, its mean has to be > 0.

Recall: CLT means that $\overline{x} \sim \mathcal{N}(\mu, \sigma^2/N)$.

$$\alpha \equiv P(\overline{x} > c \mid H_0) = P(\overline{x} > c \mid \mu = 0) = P\left(\frac{\overline{x}}{\sigma/\sqrt{N}} > \frac{c}{\sigma/\sqrt{N}}\right) = 1 - \Phi\left(\frac{c}{\sigma/\sqrt{N}}\right) = 0.05.$$

$\Longrightarrow c\sqrt{N} = \sigma\Phi^{-1}(1 - 0.05) = 4\times$`scipy.stats.norm.ppf(0.95)`$\Longrightarrow c\sqrt{N} \approx 4 \times 1.64$.

Likelihood-ratio test example

$X_i (i = 1, \cdots, N) \sim \mathcal{N}(\mu, \sigma^2)$ with $\sigma = 4$ and μ unknown. H_0: $\mu = 0$, H_A: $\mu = 4$.

Find N and LR threshold such that we are able to reject H_0 at significance $\alpha = 0.05$ and our test has power $1 - \beta = 0.95$.

For both hypotheses, $\mathscr{L}(\mu) = \prod_{i=1}^{N} \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^N \exp\left[-\frac{1}{2}\left(\frac{x_i - \mu}{\sigma}\right)^2\right] \implies \ln \mathscr{L}(\mu) = \text{const.} - \frac{1}{2\sigma^2}\sum_{i=1}^{N}(x_i - \mu)^2$

$$\implies \ln LR = -\frac{1}{2\sigma^2}\sum_{i=1}^{N}\left((x_i - \mu_2)^2 - (x_i - \mu_1)^2\right) = -\frac{1}{2\sigma^2}\sum_{i=1}^{N}\left(2x_i(\mu_1 - \mu_2) + \mu_2^2 - \mu_1^2\right)$$

Plugging in $\mu_1 = 0, \mu_2 = 4$, $\ln LR = -\frac{1}{2\sigma^2}\sum_{i=1}^{N}\left(-8x_i + 16\right) = \frac{N}{4}\left(\overline{x} - 2\right)$

In order to reject H_0, we need $\ln LR = \frac{N}{4}\left(\overline{x} - 2\right) >$ some threshold. Since N is (an unknown) constant, we need $\overline{x} >$ some threshold c (say).

this makes sense – in order to distinguish the data from noise, its mean has to be > 0.

Recall: CLT means that $\overline{x} \sim \mathcal{N}(\mu, \sigma^2/N)$.

$$\alpha \equiv P(\overline{x} > c \mid H_0) = P(\overline{x} > c \mid \mu = 0) = P\left(\frac{\overline{x}}{\sigma/\sqrt{N}} > \frac{c}{\sigma/\sqrt{N}}\right) = 1 - \Phi\left(\frac{c}{\sigma/\sqrt{N}}\right) = 0.05.$$

$\implies c\sqrt{N} = \sigma\Phi^{-1}(1 - 0.05) = 4\times$`scipy.stats.norm.ppf(0.95)`$\implies c\sqrt{N} \approx 4 \times 1.64$.

Similarly, $1 - \beta \equiv P(\overline{x} > c \mid H_A) = P(\overline{x} > c \mid \mu = 4) = 0.95 \implies \frac{c - 4}{\sigma/\sqrt{N}} = -1.64$.

$\implies N \geq 11, c \geq 2$.

Parametric tests

Tests in which either H_0 or the test statistic corresponding to H_0 assumes a distribution with associated parameters.

One-sample test: compare a parameter for a test sample to a distribution specified by H_0.
Two-sample test: compare a parameter between two test samples.

Parametric tests

Tests in which either H_0 or the test statistic corresponding to H_0 assumes a distribution with associated parameters.

One-sample test: compare a parameter for a test sample to a distribution specified by H_0.
Two-sample test: compare a parameter between two test samples.

Do they have the same mean?

- known variance or $N > 30$: use Z statistic (*Z*-test).
- unknown variance and $N < 30$: use t statistic (*t*-test).

Parametric tests

Tests in which either H_0 or the test statistic corresponding to H_0 assumes a distribution with associated parameters.

One-sample test: compare a parameter for a test sample to a distribution specified by H_0.
Two-sample test: compare a parameter between two test samples.

Do they have the same mean?

known variance or $N > 30$: use Z statistic (Z-test).

unknown variance and $N < 30$: use t statistic (t-test).

Both these tests compare data to normal distributions. There are other tests for non-normal distributions.

Parametric tests

Tests in which either H_0 or the test statistic corresponding to H_0 assumes a distribution with associated parameters.

One-sample test: compare a parameter for a test sample to a distribution specified by H_0.
Two-sample test: compare a parameter between two test samples.

Do they have the same mean?

known variance or $N > 30$: use Z statistic (Z-test).

unknown variance and $N < 30$: use t statistic (t-test).

Both these tests compare data to normal distributions. There are other tests for non-normal distributions.

Do they have the same variance?

Use the F-test.

Z- and t-test example (one sample)

A manufacturer claims that their 4M-pixel CCD detectors maintain an average of 1000 bad pixels. An inspection of 40 sample CCDs showed an average of 1200 bad pixels with a standard deviation of 500. Can the company's claim be rejected at the 5% significance level?

Z- and t-test example (one sample)

A manufacturer claims that their 4M-pixel CCD detectors maintain an average of 1000 bad pixels. An inspection of 40 sample CCDs showed an average of 1200 bad pixels with a standard deviation of 500. Can the company's claim be rejected at the 5% significance level?

H_0: $\mu = 1000$. H_A: $\mu \neq 1000$ (two-tailed test).
The standard deviation was estimated from the data, but $N > 30$, so we can use the Z-statistic.

Z- and t-test example (one sample)

A manufacturer claims that their 4M-pixel CCD detectors maintain an average of 1000 bad pixels. An inspection of 40 sample CCDs showed an average of 1200 bad pixels with a standard deviation of 500. Can the company's claim be rejected at the 5% significance level?

H_0: $\mu = 1000$. H_A: $\mu \neq 1000$ (two-tailed test).

The standard deviation was estimated from the data, but $N > 30$, so we can use the Z-statistic.

$$Z \equiv \frac{\overline{x} - \mu}{\sigma/\sqrt{N}} \approx \frac{\overline{x} - \mu}{s/\sqrt{N}} = \frac{1150 - 1000}{500/\sqrt{40}} \approx 1.898.$$

Z- and t-test example (one sample)

A manufacturer claims that their 4M-pixel CCD detectors maintain an average of 1000 bad pixels. An inspection of 40 sample CCDs showed an average of 1200 bad pixels with a standard deviation of 500. Can the company's claim be rejected at the 5% significance level?

H_0: $\mu = 1000$. H_A: $\mu \neq 1000$ (two-tailed test).

The standard deviation was estimated from the data, but $N > 30$, so we can use the Z-statistic.

$$Z \equiv \frac{\overline{x} - \mu}{\sigma/\sqrt{N}} \approx \frac{\overline{x} - \mu}{s/\sqrt{N}} = \frac{1150 - 1000}{500/\sqrt{40}} \approx 1.898.$$

The p-value is $p \equiv P(Z > 1.898) = 1 - \Phi(1.898) = 1-$ `scipy.stats.norm.cdf(1.898)` $\approx 0.03 < \alpha = 0.05$.

Therefore, the claim is rejected at the 5% significance level.

Z- and t-test example (one sample)

A manufacturer claims that their 4M-pixel CCD detectors maintain an average of 1000 bad pixels. An inspection of 40 sample CCDs showed an average of 1200 bad pixels with a standard deviation of 500. Can the company's claim be rejected at the 5% significance level?

H_0: $\mu = 1000$. H_A: $\mu \neq 1000$ (two-tailed test).
The standard deviation was estimated from the data, but $N > 30$, so we can use the Z-statistic.

$$Z \equiv \frac{\overline{x} - \mu}{\sigma/\sqrt{N}} \approx \frac{\overline{x} - \mu}{s/\sqrt{N}} = \frac{1150 - 1000}{500/\sqrt{40}} \approx 1.898.$$

The p-value is $p \equiv P(Z > 1.898) = 1 - \Phi(1.898) = 1 -$ `scipy.stats.norm.cdf(1.898)` $\approx 0.03 < \alpha = 0.05$.
Therefore, the claim is rejected at the 5% significance level.

See documentation for `statsmodels.stats.weightstats.ztest` – options and alternatives!

Z- and t-test example (one sample)

A manufacturer claims that their 4M-pixel CCD detectors maintain an average of 1000 bad pixels. An inspection of 40 sample CCDs showed an average of 1200 bad pixels with a standard deviation of 500. Can the company's claim be rejected at the 5% significance level?

H_0: $\mu = 1000$. H_A: $\mu \neq 1000$ (two-tailed test).

The standard deviation was estimated from the data, but $N > 30$, so we can use the Z-statistic.

$$Z \equiv \frac{\overline{x} - \mu}{\sigma/\sqrt{N}} \approx \frac{\overline{x} - \mu}{s/\sqrt{N}} = \frac{1150 - 1000}{500/\sqrt{40}} \approx 1.898.$$

The p-value is $p \equiv P(Z > 1.898) = 1 - \Phi(1.898) = 1-$ `scipy.stats.norm.cdf(1.898)` $\approx 0.03 < \alpha = 0.05$.

Therefore, the claim is rejected at the 5% significance level.

See documentation for `statsmodels.stats.weightstats.ztest` – options and alternatives!

Now, assume $N = 20$. We have to use the t-statistic.

Can the inspector reject the company's claim at the 5% level?

See documentation for `scipy.stats.ttest_1samp` – options and alternatives!