Data Analysis in Astronomy and Physics

Lecture 6: Hypothesis Testing

M. Röllig

Introduction to inference

Example: Gender Discrimination

48 male bank supervisors given the same personnel file, were asked to judge whether the person should be promoted. The files were identical, except for gender of applicant which was assigned randomly. 35 of the 48 supervisors voted for promotion.

Are females unfairly discriminated against?

		promotion		
		promoted	not promoted	total
gender	male	21	3	24
	female	14	10	24
	Total	35	13	48

% of males promoted= 21/24≈88% % of females promoted= 14/24≈58%

Example credits: Dr. Mine Çetinkaya-Rundel, Duke University

Example: Gender Discrimination

We have to decide between two competing claims:

null hypothesis H_0	"There is nothing going on" promotion and gender are independent, no gender discrimination, observed difference due to chance
alternative hypothesis H_A	"There is something going on" promotion and gender are dependent, there is gender discrimination

The burden of proof is on the side of H_A

Question: "Could these data plausibly have happened by chance if the null hypothesis were true?"

Answer: **yes**: fails to reject H_0
no: rejects H_0

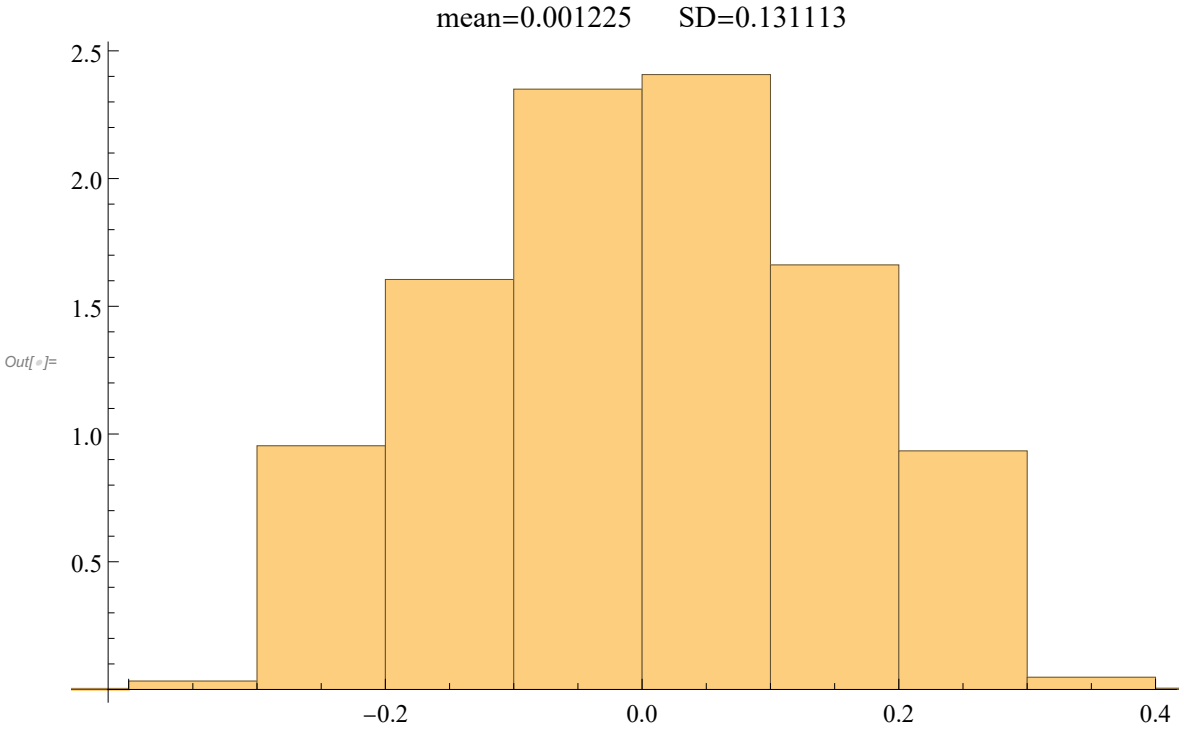
Hypothesis testing framework

- start with a null hypothesis representing the status quo
- set an alternative hypothesis representing the research question
- conduct a hypothesis test **under the assumption that the null hypothesis is true**, either via simulation or theoretical methods
  - if the test results suggest that the data do not provide convincing evidence for the alternative hypothesis, stick with the null hypothesis
  - if they do, then reject the null hypothesis in favor of the alternative

Example: Gender Discrimination - Simulation Scheme

We simulate the experiment.

Out[ ]=



Decision

If the simulation results look like the data, then the difference between the proportions of promoted files between males and females was due to chance.

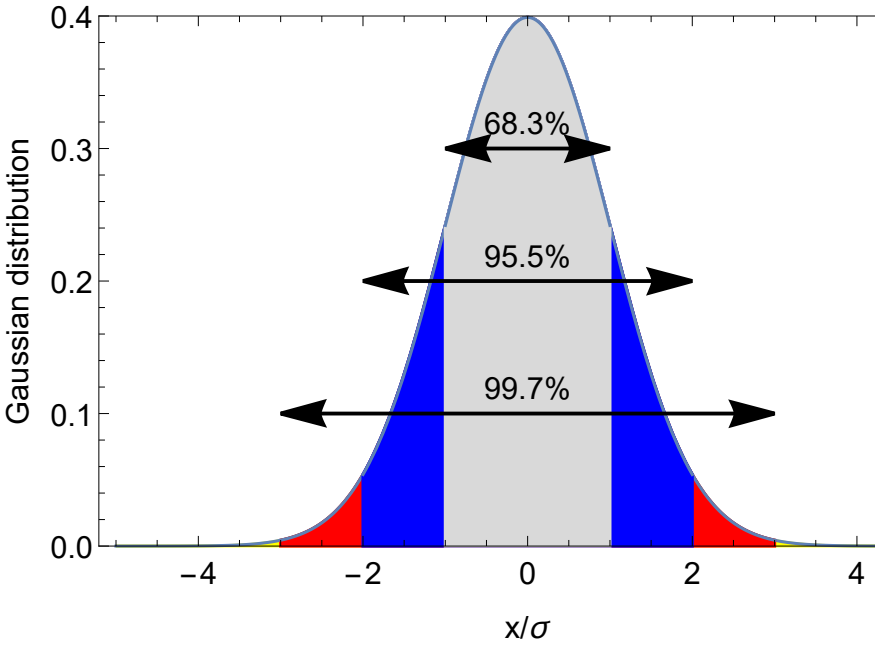
If the results from the simulations do not look like the data then the difference between the proportions of promoted files between males and females was not due to chance, but due to gender discrimination.

Confidence Interval

A plausible range for a population parameter is called confidence interval (CI).

When estimating a population parameter (point estimate) it is unlikely that we hit the exact pop. parameter. By reporting a range of plausible values it is more likely that we catch the real pop. parameter.

	Percentage area under the Gaussian curve		
x/σ	one tail	both tails	between tails
0.	50.	100.	0.
0.5	30.8538	61.7075	38.2925
1.	15.8655	31.7311	68.2689
1.5	6.68072	13.3614	86.6386
2.	2.27501	4.55003	95.45
2.5	0.620967	1.24193	98.7581
3.	0.13499	0.26998	99.73
3.5	0.0232629	0.0465258	99.9535
4.	0.00316712	0.00633425	99.9937
4.5	0.000339767	0.000679535	99.9993
5.	0.0000286652	0.0000573303	99.9999



According to the CLT a point estimate from a population sample is distributed according to

$$\overline{x} \approx \mathcal{N}\left(\text{mean} = \mu, \text{SE} = \frac{\sigma}{\sqrt{n}}\right)$$

Confidence Interval

$$\overline{x} \approx \mathcal{N}\left(\text{mean} = \mu, \text{SE} = \frac{\sigma}{\sqrt{n}}\right)$$

Providing the confidence interval:

$$\overline{x} \pm 2\,\text{SE}$$

means, we are confident, that approximately 95% of random samples will have sample means that are within two standard errors of the population mean.

± 2 SE is called the **margin of error**

Confidence Interval

Confidence interval for a population mean: Computed as the sample mean ± a margin of error (critical value corresponding to the central X% of the normal distribution times the standard error of the sampling distribution)

Out[•]//TraditionalForm=

$$\bar{x} \pm z^* \frac{s}{\sqrt{n}}$$

Conditions for CI

1. **Independence**: Sampled observations must be independent.
   - random sample/assignment
   - if sampling without replacement, n < 10% of population
2. **Sample size/skew:** Either the population distribution is normal, or if the population distribution is skewed, the sample size is large (rule of thumb: n > 30).

Confidence Interval

critical value z^*

We need to find the Z score that corresponds to our goal probability (percentage area under the Gaussian curve between two tails)

Example: what z^* corresponds to the central 95% of the Normal distribution?

```
Solve[∫_-s^s PDF[NormalDistribution[0, 1], x] dx == 0.95, s]
```

```
{{s → 1.95996}}
```

or, since the central 95% leave 2.5% of the Gaussian tails on either side we can use:

```
InverseCDF[NormalDistribution[0, 1], 0.975]
```

```
1.95996
```

Confidence Interval

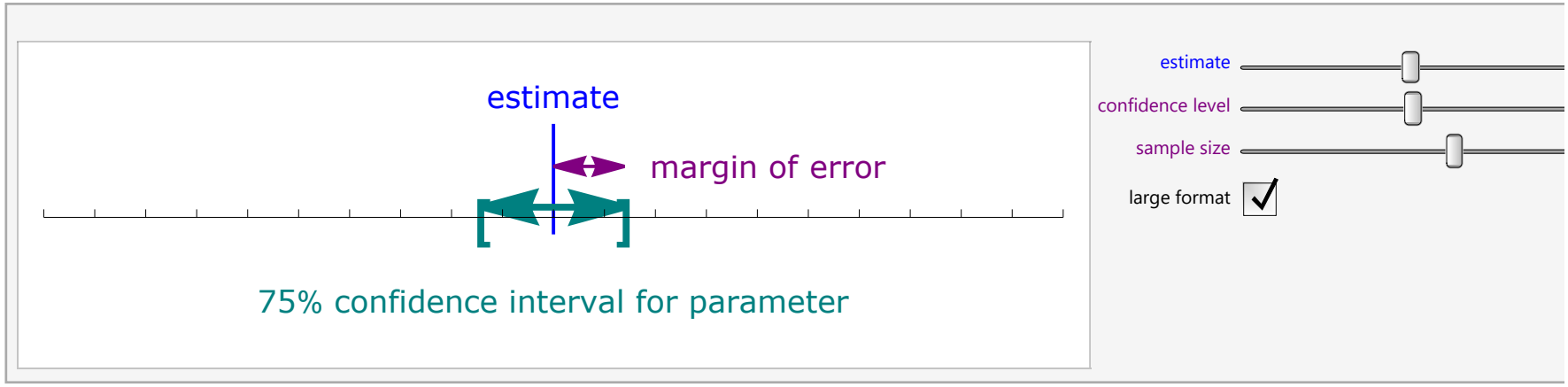
critical value z^*

similarly we can use tables to find the critical value:

	0.05	0.06	0.07	0.08	0.09
1.5	0.939429	0.94062	0.941792	0.942947	0.944083
1.6	0.950529	0.951543	0.95254	0.953521	0.954486
1.7	0.959941	0.960796	0.961636	0.962462	0.963273
1.8	0.967843	0.968557	0.969258	0.969946	0.970621
1.9	0.974412	**0.975002**	0.975581	0.976148	0.976705
2.	0.979818	0.980301	0.980774	0.981237	0.981691

a 95% CI corresponds to a margin of error of ±1.96×SE, $z^* = 1.96$

Confidence Interval



THIS NOTEBOOK IS THE SOURCE CODE FROM

"Confidence Intervals: Confidence Level, Sample Size, and Margin of Error" from the Wolfram Demonstrations Project
http://demonstrations.wolfram.com/ConfidenceIntervalsConfidenceLevelSampleSizeAndMarginOfError/

- Contributed by: Eric Schulz

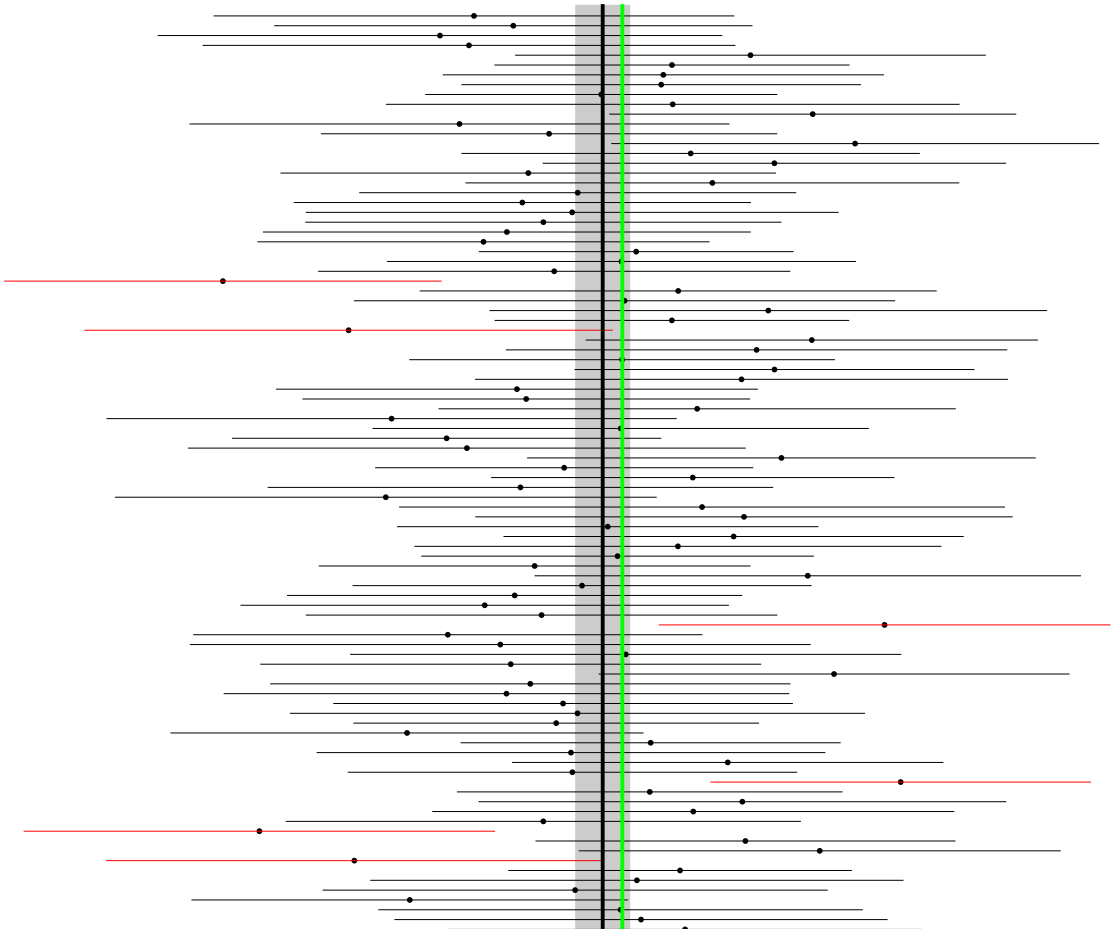
Confidence Interval

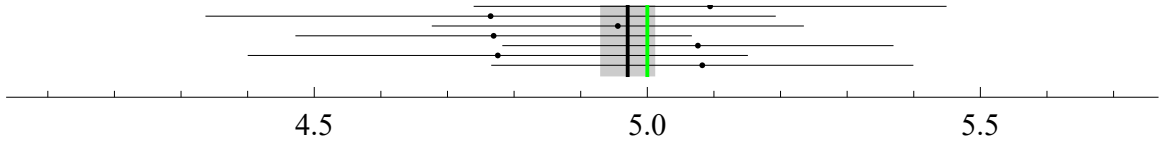
Evaluate

population: $\mathcal{N}(\mu=5, \sigma=1)$

100 samples of size 30

$\bar{x}=4.97055 \pm 0.0393134$





Avoid common misconceptions

Sample distributions

Each individual sample distribution i can be described with

$\overline{x_i}$=mean(x_i)±z^*×SD$_i$/$\sqrt{\text{sample size}}$

$\overline{x_i}$=mean(x_i)±z^*×SE$_i$

Interpretation:

When sampling 100 samples of size 30 from the population, there is a 95% chance that the population mean μ is part of the CI mean(x_i)±z^*×SE$_i$

The sampling distribution is different!

μ=mean($\overline{x_i}$)±z^*×SD($\overline{x_i}$)/$\sqrt{\#\text{ of samples}}$

μ=mean($\overline{x_i}$)±z^*×SE

Interpretation:

When sampling 100 samples of size 30 from the population, there is a 95% chance that the population mean μ is part of the CI mean($\overline{x_i}$)±z^*×SE

Quiz

Based on a sample of 20 asteroids from Britt et al. (2002) we compute a 95% confidence interval for the average bulk density of asteroids of 2.26±0.56 g/cm^3.

Determine if each of the following statements are true or false.

95% of all asteroids have a bulk density between 1.67 and 2.82 g/cm^3.

95% of random samples of 20 asteroids will yield confidence intervals that contain the true average bulk density of asteroids.

95% of the time, the true average bulk density of asteroids is between 1.67 and 2.82 g/cm^3.

We are 95% confident that asteroids in this sample have an bulk density on average of 1.67 to 2.82 g/cm^3.

Confidence Interval

Given a target margin of error, confidence level, and information on the variability of the sample (or the population), we can determine the required sample size to achieve the desired margin of error.

Out[ ]=

$$\text{ME} = z^* \frac{s}{\sqrt{n}} \quad \Longrightarrow \quad n = \left(\frac{z^* \, s}{\text{ME}} \right)^2$$

Example:

How many additional asteroids do we need to measure in order to halve the margin of error on the average bulk density?

Out[•]= $n = \left(\frac{z^* s}{ME}\right)^2, \quad ME' = \frac{ME}{2} \quad \Longrightarrow \quad n' = \left(\frac{z^* s}{ME'}\right)^2 = \left(\frac{z^* s}{\frac{1}{2}ME}\right)^2 = 4\,n$

We need a sample of 80 asteroids.

Suppose the researchers think a 99% confidence level would be more appropriate for this interval. Will this new interval be narrower or wider than the 95% confidence interval?

Hypotheses

null H_0	Often either a skeptical perspective or a claim to be tested =
alternative H_A	Represents an alternative claim under consideration and is often ≠, >, <, ... represented by a range of possible parameter values.

The conservative/sceptic will not abandon the H_0 unless the evidence in favor of the H_A is so strong that she/he rejects H_0 in favor of H_A

Hypothesis testing for a mean

Example

Earlier we computed a 95% CI for the average bulk density of asteroids to be between 1.67 and 2.82 g/cm^3.

Based on this confidence interval do these data support the hypothesis, that asteroids on average have a density above 2 g/cm^3?

$H_0 : \mu_\rho = 2\ g/\text{cm}^3$ Asteroids have an average density of $2\ g/\text{cm}^3$.

$H_A: \mu_\rho > 2\ g/\text{cm}^3$ Asteroids have on average a density larger than $2\ g/\text{cm}^3$.

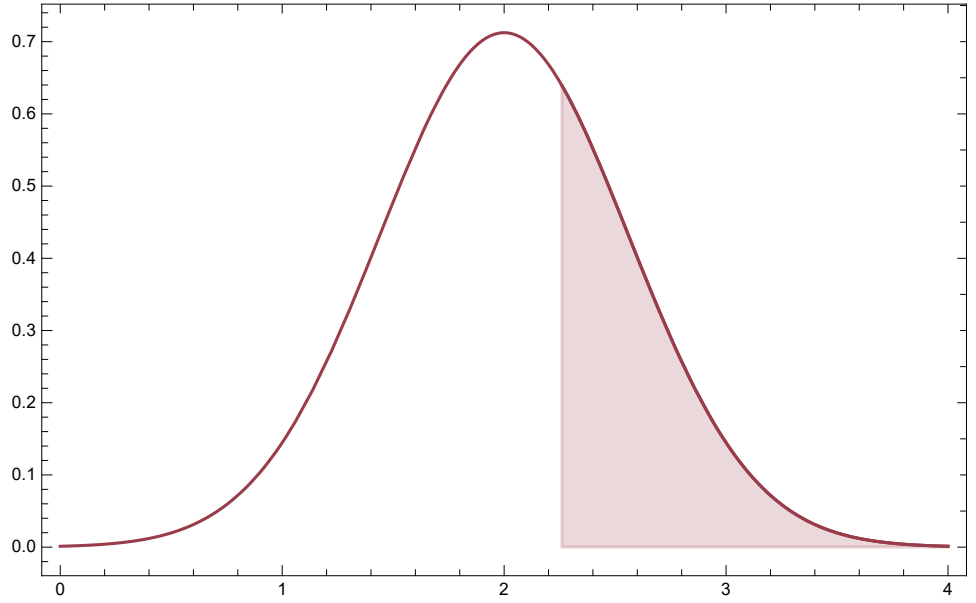
Hypothesis testing is always about population parameters (here μ_ρ)! Never about sample parameters (e.g. $\overline{\rho}$)!

Hypothesis testing for a mean

p-Value

Out[•]//TraditionalForm=

$$p = \mathcal{P}_r(\text{observed or more extreme outcome} \mid H_0 \text{ true})$$



$$p = \mathcal{P}_r(\bar{\rho} > 2.26 \mid H_0{:}\mu_\rho{=}2)$$

$$\bar{\rho}_{\text{asteroids}} \approx \mathcal{N}(\mu_\rho{=}2, \text{SE}{=}0.56)$$

Out[•]=

n=20

$\bar{\rho}_{\text{asteroid}}$=2.26 g/cm^3

sample SD=1.28 g/cm^3

$$\text{SE} = \frac{\text{SD}}{\sqrt{n}} = \frac{1.28}{\sqrt{20}} = 0.56$$

test statistic: $Z = \frac{\text{obs}-\text{pop}}{\text{SE}} = \frac{2.26-2}{0.56} = 0.46$

p-value = $\mathcal{P}_r(Z > 0.46) = 0.32$

```
1 - CDF[NormalDistribution[], 0.46]
```

```
0.322758
```

Hypothesis testing for a mean

Decision

We used the test statistic to calculate the **p-value**, the probability of observing data at least as f**avorable to the alternative hypothesis** as our current data set, **if the null hypothesis was true**!

If the **p-value is low** (lower than the significance level α, usually set to 5%) then it would be very unlikely to observe the data (as we did) if the null hypothesis were true. Hence **we reject H_0**.

If the **p-value is high** (higher than our significance level α) then it is not unlikely to observe just the data that we observed if the null hypothesis were true. Hence **we do not reject H_0.**

Hypothesis testing for a mean

Out[•]//TraditionalForm=

$$p = \mathcal{P}(\text{observed or more extreme outcome} \mid H_0 \text{ true})$$

$$p = \mathcal{P}(\bar{\rho} > 2.26 \mid H_0{:}\mu_\rho{=}2)$$

$$\bar{\rho}_{\text{asteroids}} \approx \mathcal{N}(\mu_\rho{=}2, \text{SE}{=}0.56)$$

Out[•]=

n=20

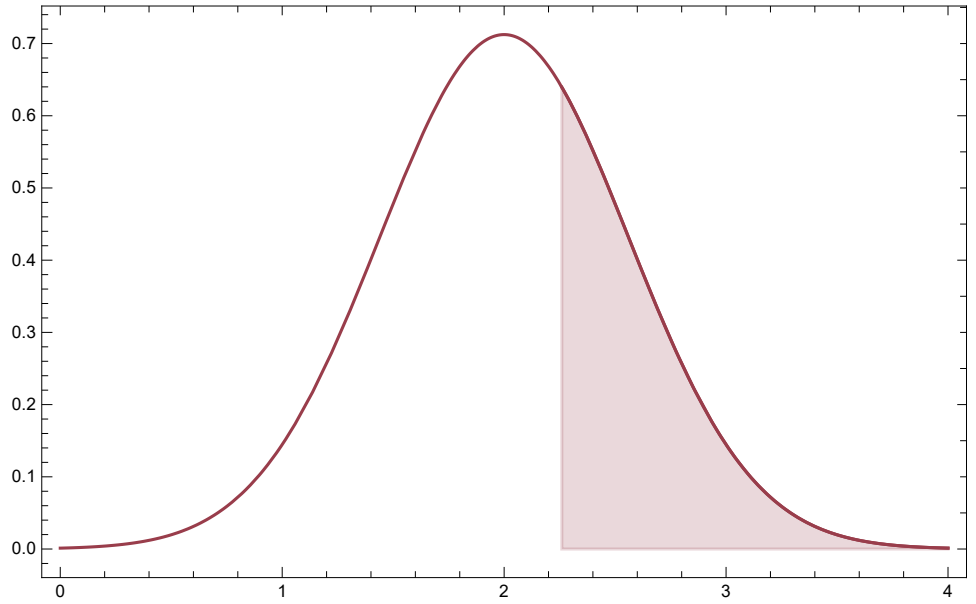
$\bar{\rho}_{\text{asteroid}}$=2.26 g/cm^3

sample SD=1.28 g/cm^3

$$\text{SE} = \frac{\text{SD}}{\sqrt{n}} = \frac{1.28}{\sqrt{20}} = 0.56$$

$$\text{Z} = \frac{\text{obs}-\text{pop}}{\text{SE}} = \frac{2.26-2}{0.56} = 0.46$$

p−value = $\mathcal{P}$(Z>0.46)=0.32



The p-value is pretty high (>30%). This means there is a larger than 30% chance to find the data that we observed if the null hypothesis were true, i.e. asteroids have a mean density of 2 g/cm^3. We do not reject H_0.

Observing a sample of asteroids with mean density 2.26 g/cm^3 is likely to happen simply by chance.

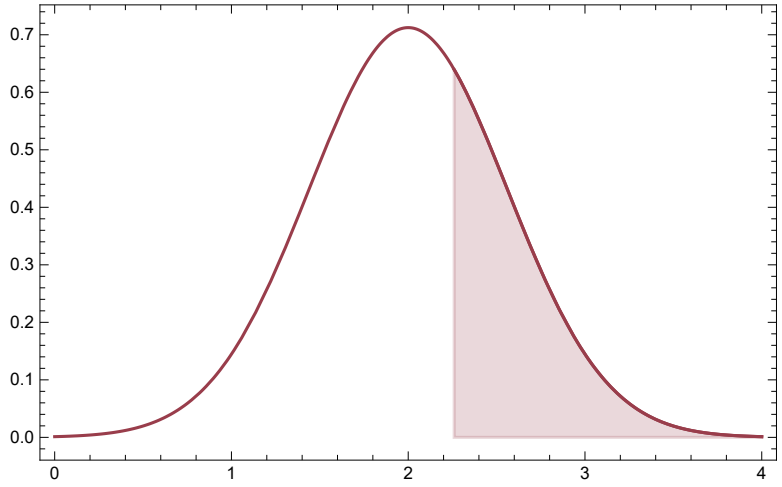
The data do not provide convincing evidence that asteroids on average have densities above 2 g/cm^3.

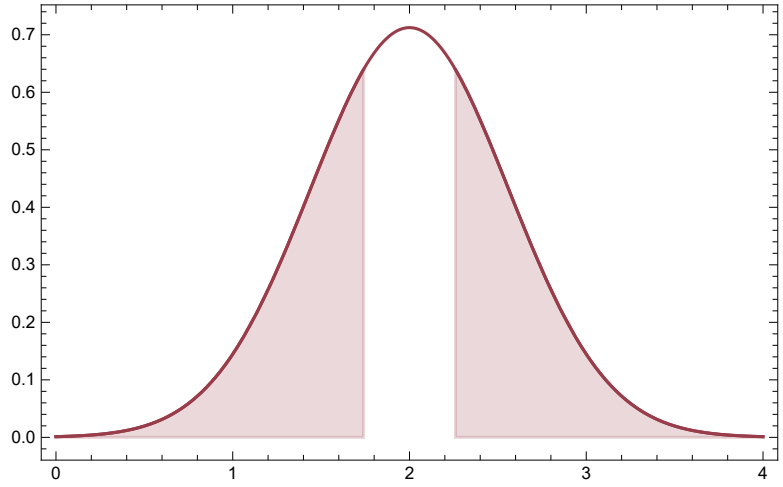
The difference between the null value of 2 g/cm^3 and the observed sample mean 2.26 g/cm^3 is due to **chance** or **sampling variability**.

Two-sided vs. one-sided tests

When looking at the divergence from null in any direction (instead of one direction) we perform a two-sided hypothesis test. The definition of the p-value remains the same, but we have to make sure to cover all directions.

Out[ ]=





Out[ ]//TraditionalForm=

$$\mathcal{P}_r\left(\bar{\rho} > 2.26 \text{ OR } \bar{\rho} < 1.74 \,\middle|\, H_0{:}\mu_\rho{=}2\right)$$

$$\begin{aligned} p - \text{value} &= \mathcal{P}_r(Z > 0.46) - \mathcal{P}_r(Z < -0.46) \\ &= 2 \times 0.32 \\ &= 0.64 \end{aligned}$$

Hypothesis testing for a single mean

1. Set the hypotheses: $H_0 : \mu = null\ value$
   $H_A : \mu <$ or $>$ or $\neq$ $null\ value$
2. Calculate the point estimate: $\bar{x}$
3. Check conditions:
   3.1. **Independence**: Sampled observations must be independent (random sample/assignment & if sampling without replacement, n < 10% of population)
   3.2. **Sample size/skew:** n≥30, larger if the population distribution is very skewed.
4. Draw sampling distribution, shade p-value, calculate test statistic $Z = \frac{\bar{x}-\mu}{SE}$, $SE = \frac{s}{\sqrt{n}}$
5. Make a decision, and interpret it in context of the research question:
   - **if p-value <α, reject H_0;** the data provide convincing evidence for H_A
   - **if p-value >α, fail to reject H_0;** the data **do not** provide convincing evidence for H_A

Example

Researchers investigating characteristics of gifted children collected data from schools in a large city on a random sample of thirty-six children who were identified as gifted children soon after they reached the age of four. In this study, along with variables on the children, the researchers also collected data on their mothers' IQ scores.

n=36, min=101, mean=118.2, sd=6.5, max=131

Perform a hypothesis test to evaluate if these data provide convincing evidence of a difference between the average IQ score of mothers of gifted children and the average IQ score for the population at large, which is 100. Use a significance level of 0.01.

1. Set the hypothesis μ=average IQ score of mothers of gifted children
   $H_0 : \mu = 100$ $H_A : \mu \neq 100$
2. Calculate the point estimate: $\bar{x} = 118.2$
3. Check conditions:
   1. random & 35<10% of all gifted children → independent
   2. n>30 & sample not skewed → nearly normal sampling distribution

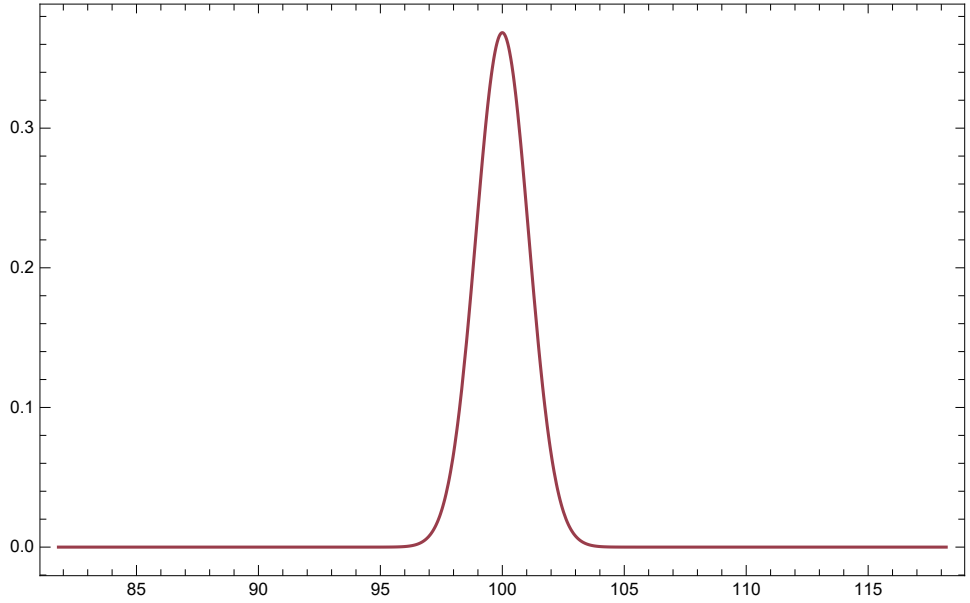
Example

$H_0 : \mu = 100$ $\quad\quad$ $\bar{x} = 118.2$

$H_A : \mu \neq 100$ $\quad\quad$ $\bar{x} \approx \mathcal{N}\left(\mu = 100, \text{SE} = \frac{s}{\sqrt{n}} = \frac{6.5}{\sqrt{36}} \approx 1.083\right)$

4. Draw sampling distribution, shade p-value, calculate test statistic

Out[•]=



$Z = \frac{118.2-100}{1.083} = 16.8$

p-value ≈ 0

5. Make a decision:
p-value is very low ⟶ strong evidence against the null.

We reject the null hypothesis and conclude that the data provide convincing evidence of a difference between the average IQ score of mothers of gifted children and the average IQ score for the population at large.

HT for paired numerical data

If we want to compare how two numerical variables of a data set differ we perform a hypothesis test for paired data. “Paired” means in this context that each case in our data set contains the two variables of interest.

Example:

200 observations were randomly sampled from the stellar survey. The same stars are given with brightnesses in the blue (B) and visual (V) bands. Are B and V different from of each other?

	ID	B	V
1	70	5.7	5.2
2	86	4.4	3.3
3	141	6.3	4.4
4	172	4.7	5.2
...	...	...	...
200	137	6.3	6.5

- When two sets of observations have this correspondence (not independent) they are said to be paired.
- To analyze paired data, it is often useful to look at the difference of the pair:
  diff=B-V
- It is important to always subtract using a consistent order.

Example:

200 observations were randomly sampled from the stellar survey. The same stars are given with brightnesses in the blue (B) and visual (V) bands. Are B and V different from of each other?

	ID	B	V	B−V
1	70	5.7	5.2	**0.5**
2	86	4.4	3.3	**1.1**
3	141	6.3	4.4	**1.9**
4	172	4.7	5.2	**-0.5**
…	…	…	…	**…**
200	137	6.3	6.5	**-0.2**

- When two sets of observations have this correspondence (not independent) they are said to be paired.
- To analyze paired data, it is often useful to look at the difference of the pair:
  diff=B-V
- It is important to always subtract using a consistent order.

HT for paired numerical data

parameter of interest

Average difference B-V between the blue and visual magnitudes of **all** stars

$$\mu_{\text{diff}} = \mu_{B-V}$$

point estimate

Average difference B-V between the blue and visual magnitudes of **sampled** stars

$$\overline{x}_{\text{diff}} = \overline{x}_{B-V}$$

HT for paired numerical data

Example:

If there was no difference in the B and V magnitudes, what would we expect the average difference to be?

	ID	B	V	B-V
1	70	5.7	5.2	**0.5**
2	86	4.4	3.3	**1.1**
3	141	6.3	4.4	**1.9**
4	172	4.7	5.2	**-0.5**
…	…	…	…	**…**
200	137	6.3	6.5	**-0.2**

$\bar{x}_{diff} = -0.545$

$s_{diff} = 8.887$

$n_{diff} = 200$

hypotheses for paired means

$H_0 : \mu_{diff} = 0$ There is no difference between the B and V magnitudes

$H_A : \mu_{diff} \neq 0$ There is a difference between the B and V magnitudes

This way we reduced the problem to the already covered case of testing one numerical mean!

Hypothesis testing for a difference between paired means

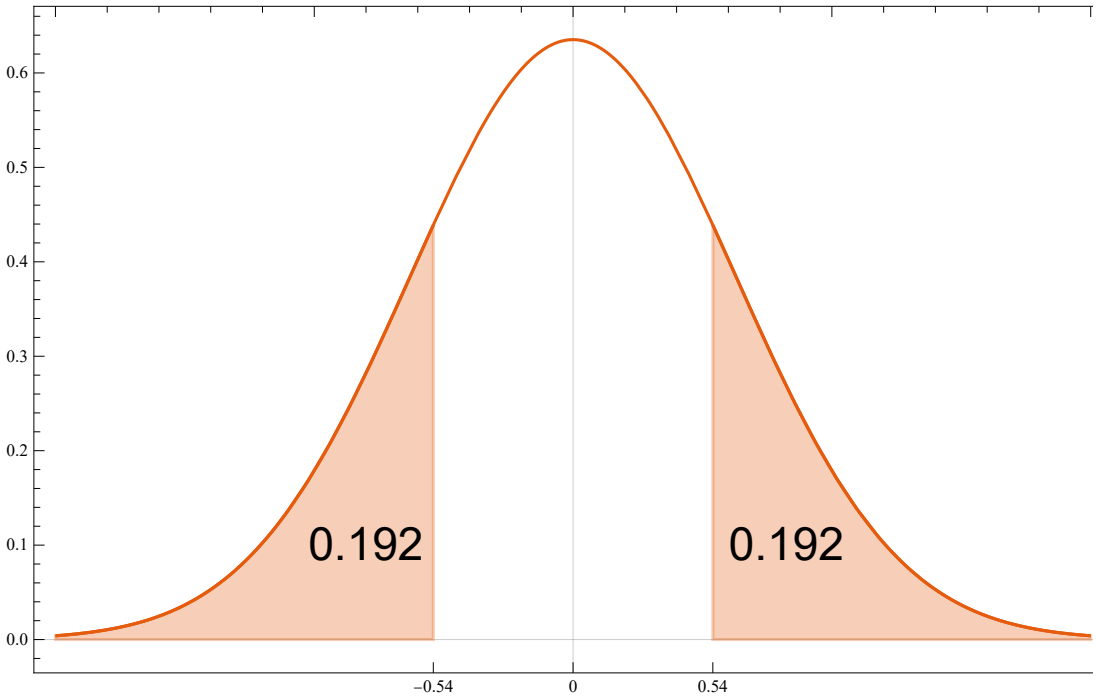
1. Set the hypotheses: $H_0 : \mu_{diff} = null\ value$
   $H_A : \mu_{diff} <$ or $>$ or $\neq$ $null\ value$
2. Calculate the point estimate: $\bar{x}_{diff}$
3. Check conditions:
   - 3.1. **Independence**: Sampled observations must be independent (random sample/assignment & if sampling without replacement, n_{diff} < 10% of population)
   - 3.2. **Sample size/skew:** $n_{diff} \geq 30$, larger if the population distribution is very skewed.
4. Draw sampling distribution, shade p-value, calculate test statistic $Z = \frac{\bar{x}_{diff} - \mu_{diff}}{SE_{\bar{x}_{diff}}}$, $SE = \frac{s_{diff}}{\sqrt{n_{diff}}}$
5. Make a decision, and interpret it in context of the research question:
   - **if p-value <α, reject H_0;** the data provide convincing evidence for H_A
   - **if p-value >α, fail to reject H_0;** the data **do not** provide convincing evidence for H_A

HT for paired numerical data

Example:

$H_0: \mu_{diff} = 0$
$H_A: \mu_{diff} \neq 0$
$\bar{x}_{diff} = -0.545$
$s_{diff} = 8.887$
$n_{diff} = 20$

$$\bar{x}_{diff} \approx \mathcal{N}(\text{mean}=0, \text{SE}=\frac{8.887}{\sqrt{200}} \approx 0.628)$$



Quiz

Which of the following is the correct interpretation of the p-value from the example?

1. Probability that the average B and V magnitudes are equal.
2. Probability that the average B and V magnitudes are different
3. Probability of obtaining a random sample of 200 stars where the average difference between the B and V magnitudes is at least 0.545 (in either direction), if in fact the true average difference between the magnitudes is 0.
4. Probability of incorrectly rejecting the null hypothesis if in fact the null hypothesis is true

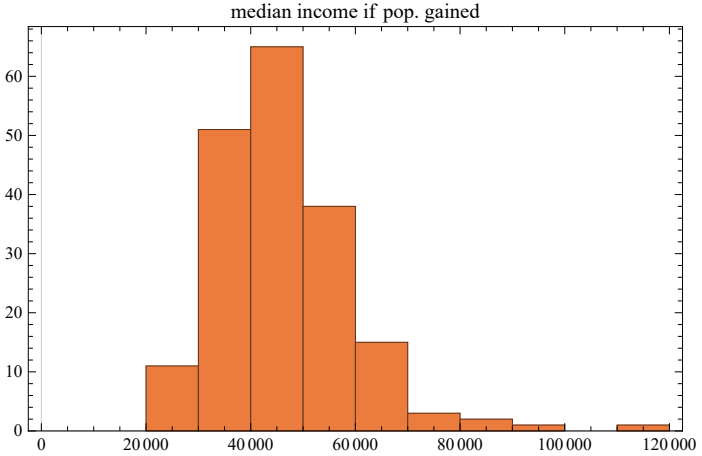
HT for paired numerical data

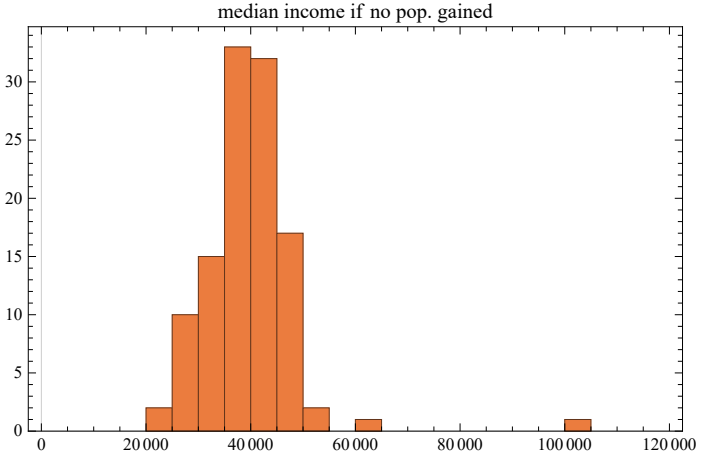
Summary

- paired data (2 variables) $\longrightarrow$ differences (1 variable)
- most often: $H_0 : \mu_{\text{diff}} = 0$
- e.g. same individuals, but comparison between multiple measurements (pre-post. repeated, time dependent, ...)
- e.g. different (but dependent) individuals, e.g twins, partners, ...

HT for two independent means

Using again the `county.txt` data set from problem set 2 we demonstrate how to test for unpaired data. Is there a difference in median income per household between counties which gained population between 2000 and 2010 and those that did not gain population? (we took a sample of 300 counties)





⋮

	$\bar{x}$	s	n
gained	46 352.4	12 829.4	187
not gained	39 640.2	8892.9	113

HT for two independent means

parameter of interest

Average difference in median income per household between counties which gained population between 2000 and 2010 and those that did not gain population for **all** households:

$$\mu_{\text{gained}} - \mu_{\text{not gained}}$$

point estimate

Average difference in median income per household between counties which gained population between 2000 and 2010 and those that did not gain population for **sampled** households

$$\overline{x}_{\text{gained}} - \overline{x}_{\text{not gained}}$$

HT for two independent means

Estimating the difference between independent means

point estimate ± margin of error

Out[•]//TraditionalForm=

$$(\bar{x}_1 - \bar{x}_2) \pm z^* \, \mathrm{SE}_{(\bar{x}_1 - \bar{x}_2)}$$

Out[•]=

Standard error of difference between two independent means:

$$\mathrm{SE}_{(\bar{x}_1 - \bar{x}_2)} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Conditions for inference for comparing two independent means

1. **Independence**:
   - **within groups:** sampled observations must be independent
     - random sample/assignment
     - if sampling without replacement, n < 10% of population)
     - **between groups:** the two groups must be independent of each other (unpaired)
2. **Sample size/skew:** Each sample size must be at least 30 ($n_1 \geq 30$ and $n_2 \geq 30$), larger if the population distribution is very skewed.

Example

Assuming a 95% confidence interval, how much higher is the average median household income in counties which gained population between 2000 and 2010 compared to those counties that did not gain population?

$(\overline{x}_{\text{gained}} - \overline{x}_{\text{not gained}}) \pm z^* \text{ SE} =$

$= (46\,685.4 - 39\,724.4) \pm 1.96 \sqrt{\frac{12\,407.3^2}{187} + \frac{8892.9^2}{113}}$

$= 6961.03 \pm 1.96 \times 1234.13$

$= 6961.03 \pm 2418.89$

$= (4542.14, 9379.92)$

HT for two independent means

- null hypothesis: no difference $H_0 : \mu_1 - \mu_2 = 0$
- alternative hypothesis: some difference $H_A : \mu_1 - \mu_2 \neq 0$
- same conditions and SE as the confidence interval

Example - test

Test whether the data provide convincing evidence, that the average median household income in counties which gained population between 2000 and 2010 is larger than that of counties that did not gain population?

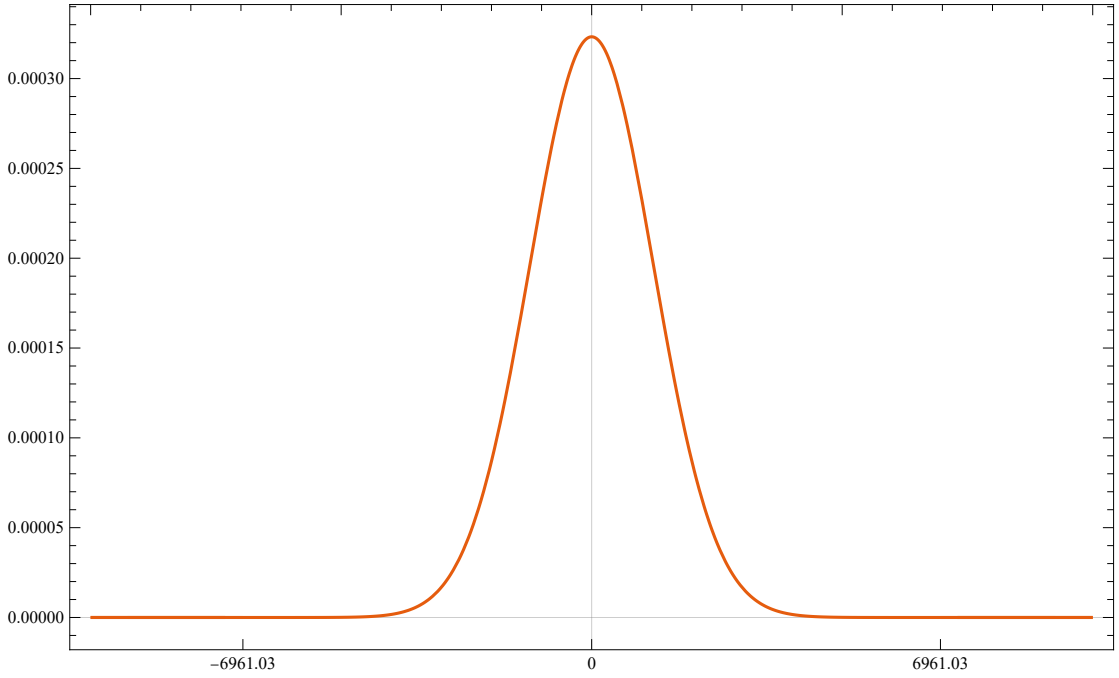
$H_0: \mu_{\text{gained}} - \mu_{\text{not gained}} = 0$

$H_A: \mu_{\text{gained}} - \mu_{\text{not gained}} \neq 0$

$\bar{x}_{\text{gained}} - \bar{x}_{\text{not gained}} = 6961.03$

$SE = 1234.128$

$\bar{x}_{\text{diff}} \approx \mathcal{N}(\text{mean}=0, SE=1234.128)$



Example - interpret the p-value

$$\text{p-value} = \mathcal{P}_r(\text{observed or more extreme statistic} \mid H_0 \text{ true})$$

If there is no difference between the median household income of counties which gained population compared to those which did, then there is a 0% chance (p-value) of obtaining random samples of 187 counties which gained population and 113 counties which did not where the average difference between their median household income is at least $6961.03.

Bootstrapping

So far we used CLT based methods to infer population information from our sampled data.

What if the CLT does not apply:

- sample size is $\lesssim 30$
- the underlying population has a strongly skewed distribution
- the statistic of interest does not obey the CLT (e.g. the median)

Bootstrapping

- Bootstrapping is an alternative approach to **construct confidence intervals**
- Here the impossible task is estimating a population parameter, and we will use only the sample data to accomplish it.

Example

We have a radio sample of 20 flux densities observed extra-galactically (numbers made up).

In[•]:=
```
radio = {7.75, 6.26, 7.33, 9.29, 8.95, 7.49, 10.20, 13.49, 5.99, 11.43, 12.09, 14.95, 8.79, 9.75, 10.76, 12.82, 6.65, 7.05, 7.99, 5.0};
Median[radio]
```

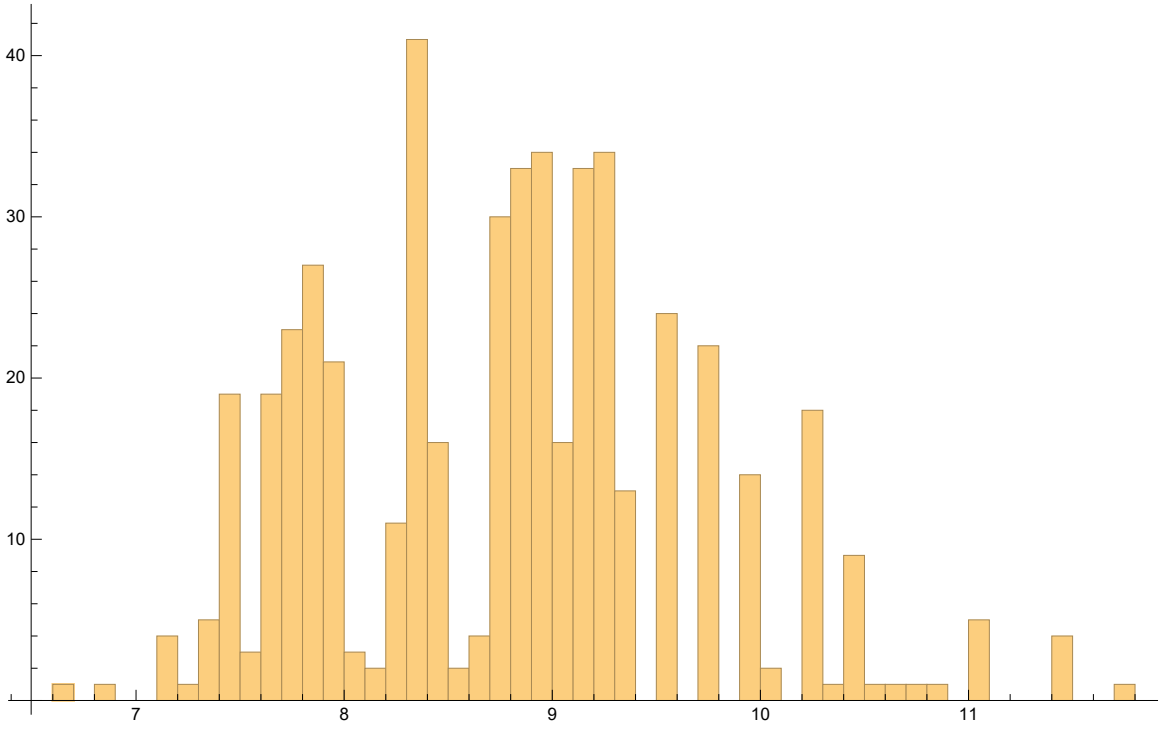
Out[•]=
```
8.87
```

Bootstrapping

1. take a bootstrap sample : this is a random sample **taken with replacement** from the original sample, of the **same size n** as the original data.
2. calculate the bootstrap statistics: a statistic such as mean, median, proportion, mode, etc. computed on the bootstrap sample.
3. repeat steps 1) and 2) many (N) times to create a bootstrap distribution - a distribution of bootstrap statistics

Example

Let's take 500 bootstrap samples and compute their medians. $\overline{\text{median}}_{\text{boot}}$ is the mean of all bootstrap medians. SE_{boot} is the standard deviation of the bootstrap medians.




```
Sample bootstrap sample:
{9.29, 7.05, 6.26, 9.29, 7.33, 5., 5., 6.65, 12.09, 12.82, 8.79, 6.65, 7.05, 8.79, 9.29, 5.99, 8.79, 7.99, 12.82, 6.65}
median_boot ± z* SE_boot = 8.81373 ± 1.96 × 0.88197 =
 = {7.08507, 10.5424}
Quantile method: {7.41, 10.69}
```

Bootstrapping

- Not as rigid conditions as CLT based methods.
- A representative sample is required for generalizability. If the sample is biased, the estimates resulting from this sample will also be biased.
- Asymptotic theory shows that the sampling distribution can be well-approximated by generating $N \simeq n \ln(n)^2$ bootstrap resamples (Babu & Singh 1983).

  E.g.: sample size n=10 ⇒ N = 53.019, n=50 ⇒ N = 765.196, n=1000 ⇒ N = 47 717.1

Student t-distribution

Conditions for inference were so far: large sample size n and population distribution is not extremely skewed - why?

Because then,

- the sampling distribution of the mean is nearly normal
- the estimate of the standard error is reliable: SE = $\frac{s}{\sqrt{n}}$

This is a consequence of the CLT. The standard error estimate is reliable if the sample size is large and because we trust s as a good estimate for σ!

What if our sample size is small and we don't know σ? Then, $\frac{s}{\sqrt{n}}$ is a bad estimator for σ.

If n is small and σ is not known, then use the t-distribution instead of the standard normal distribution to correct the uncertainty in the estimate of the standard error.

Student t-distribution

PDF - probability density function

Out[•]=

$$f(t;\nu) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\sqrt{\nu\,\pi}}\left(1+\frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}} \quad \text{or} \quad f(t;\nu) = \frac{1}{B\left(\frac{1}{2},\frac{\nu}{2}\right)\sqrt{\nu}}\left(1+\frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}}$$

where B is the Beta function and *v* is the **degree of freedom**.

For *v* even,

Out[•]//TraditionalForm=

$$\frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\sqrt{\nu\,\pi}} = \frac{(\nu-1)(\nu-3)\ldots 5\cdot 3}{2\sqrt{\nu}\,(\nu-2)(\nu-4)\ldots 4\cdot 2}$$

For *v* odd,

Out[•]//TraditionalForm=

$$\frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\sqrt{\nu\,\pi}} = \frac{(\nu-1)(\nu-3)\ldots 4\cdot 2}{\pi\sqrt{\nu}\,(\nu-2)(\nu-4)\ldots 5\cdot 3}$$

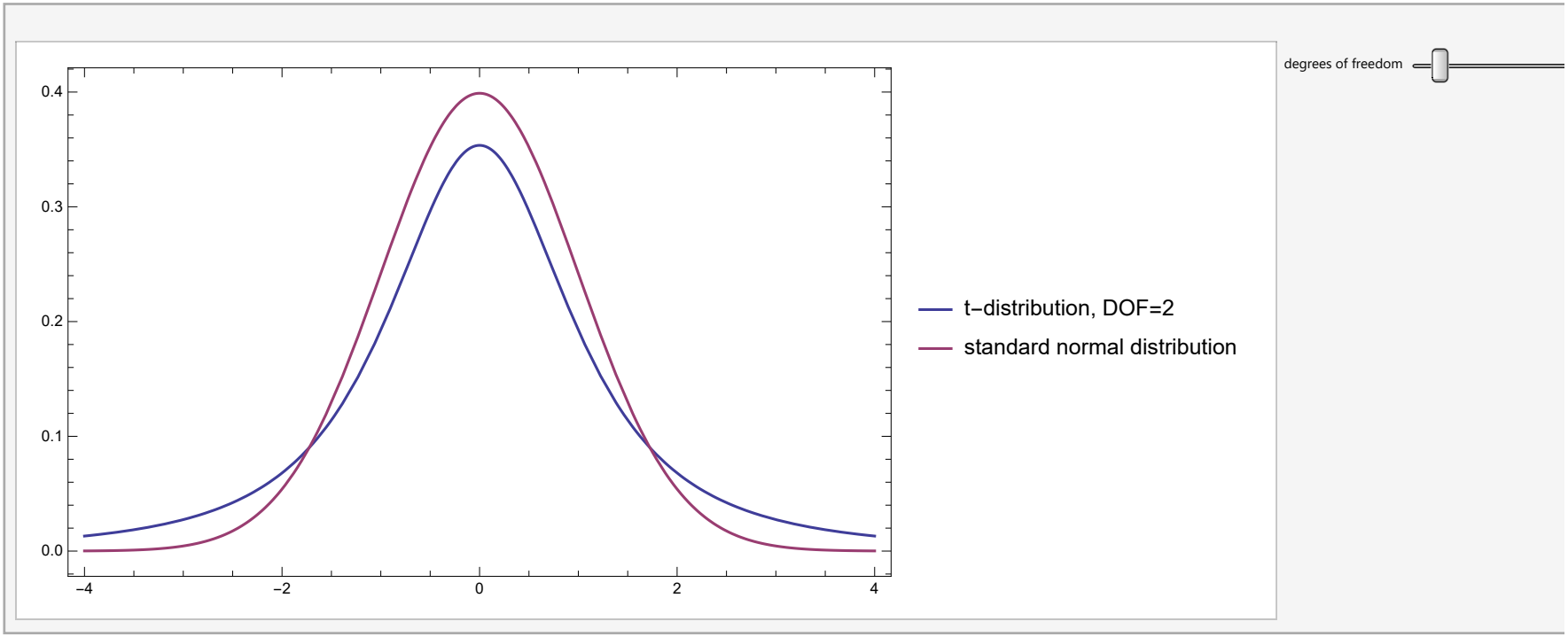
Remember, that

Out[•]=

$$\Gamma(z) = \int_0^\infty t^{z-1}\exp(-t)\,dt \quad /;\ \mathrm{Re}(z) > 0 \qquad \text{and} \qquad B(x,y) = \frac{\Gamma(x)\,\Gamma(y)}{\Gamma(x+y)} = \int_0^1 t^{x-1}(1-t)^{y-1}\,dt \quad \text{for } \mathrm{Re}(x), \mathrm{Re}(y) > 0$$

Student t-distribution

PDF - probability density function



The t-distribution has broader wings than $\mathcal{N}(0,1)$, which leads to larger p-values for a given score statistic.

Student t-distribution

Example

The Acme Chain Company claims that their chains have an average breaking strength of 20,000 kg, with a standard deviation of 1750 kg. Suppose a customer tests randomly-selected chains. What is the probability that the average breaking strength in the test will be no more than 19,800 kg?

Compute the probability for two different case: 200 test, and 14 tests.

n=200 ≳ 30 and σ is known ⇒ CLT is valid, we can use the z-score

$$Z = \frac{19\,800 - 20\,000}{1750./\sqrt{200}} = -1.61624$$

n=14 < 30 and σ is known ⇒ CLT is NOT valid, we use the t-distribution

$$T = \frac{19\,800 - 20\,000}{1750./\sqrt{14}} = -0.427618$$

Example

n=200 ≳ 30 and σ is known ⇒ CLT is valid, we can use the z-score $Z = \frac{19\,800-20\,000}{1750./\sqrt{200}} = -1.61624$

Out[•]=

$H_0: \mu = 20000$

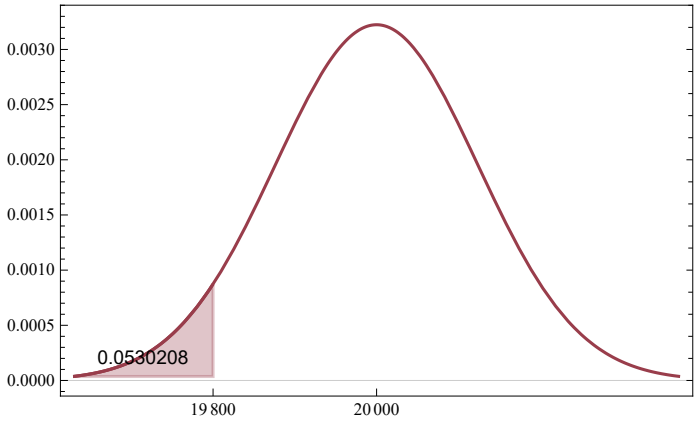
$H_A: \mu < 20000$

$\bar{x} = 19800$

s=1750

n=200

$\bar{x}_{\text{diff}} \approx \mathcal{N}(\text{mean}=20\,000;\ \text{SE}=\frac{1750}{\sqrt{200}} \approx 123.744)$



$Z = \frac{19\,800-20\,000}{\frac{1750}{\sqrt{200}}} = -1.61624$

p−value=0.0530208

Example

n=14 < 30 and σ is known ⇒ CLT is NOT valid, we use the t-distribution, $T = \frac{19\,800-20\,000}{1750./\sqrt{14}} = 0.427618$

Out[ ]=

H_0:μ=20000

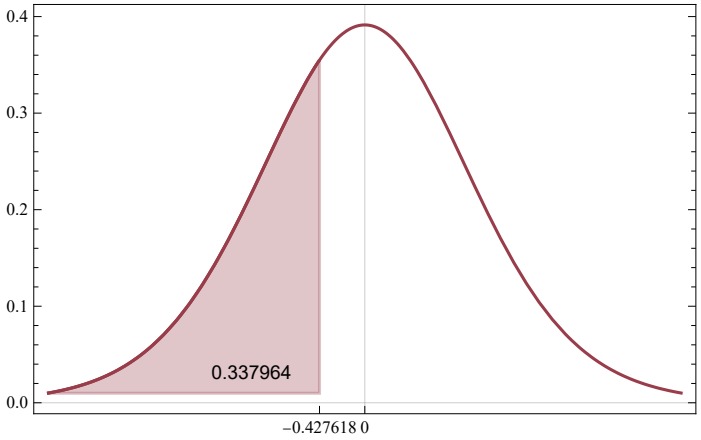
H_A:μ<20000

$\overline{x}$=19800

SE=125 $\sqrt{14}$

n=14

$\overline{x}_{\text{diff}} \approx f_{\text{Student}-t}(\text{dof}=13)$



$T=\frac{19\,800-20\,000}{\frac{125\,\text{Sqrt}[14]}{\sqrt{14}}}=-0.427618$

p−value=0.337964

Student t-distribution

How does the p-value for a given score depends on whether we use the Normal distribution or the Student's t-distribution`?

$\mathcal{P}_r(|Z|>2)=0.04550026389635842$

$\mathcal{P}_r(|t_{DOF=50}|>2)=0.050947068737692947$

$\mathcal{P}_r(|t_{DOF=10}|>2)=0.0733880347707403$

We see, that depending on what distribution we use, we might reject or fail to reject H_0!

Inference for a small sample mean

Example: Playing a computer game during lunch affects fullness,memory for lunch, and later snack intake

Sample: 22 men, 22 women

Study: 2 randomized groups, one playing Solitaire during lunch, the other eating without distraction

biscuit intake	$\bar{x}$	s	n
solitaire	52.1 g	45.1 g	22
no distraction	27.1 g	26.4 g	22

Estimating the mean for a small sample

point estimate ± margin of error

$$\bar{x} \pm t^*_{df}\, SE_{\bar{x}}$$

$$\bar{x} \pm t^*_{df}\, \frac{s}{\sqrt{n}}$$

$$\bar{x} \pm t^*_{n-1}\, \frac{s}{\sqrt{n}}$$

degrees of freedom for t statistic for inference on one sample mean

Out[•]//TraditionalForm=

$$df = n - 1$$

Determine the critical t score

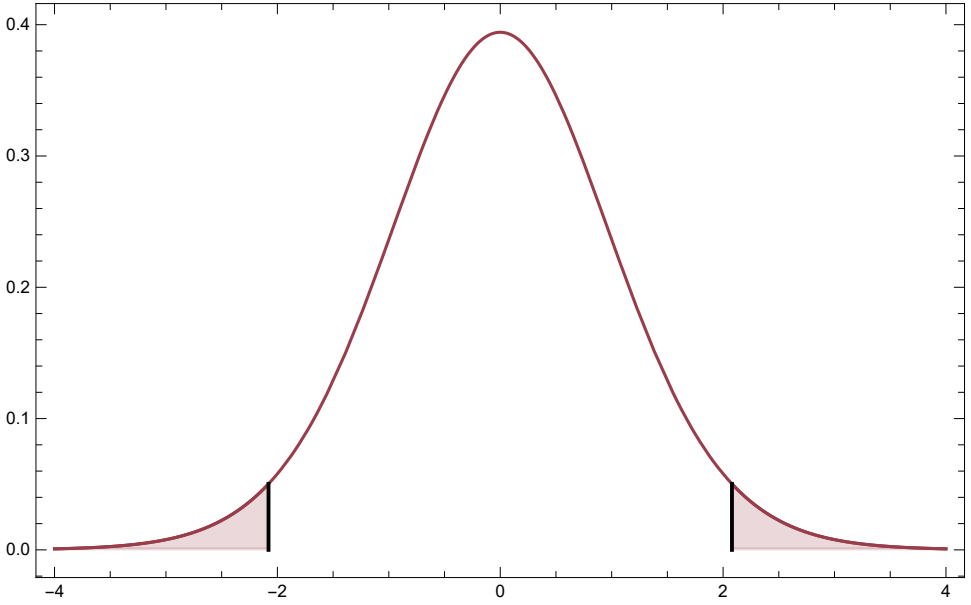
1. determine df df = $n - 1 = 22 - 1 = 21$
use a table, us R: `qt(0.025, df = 21)`, *Mathematica*:

```
InverseCDF[StudentTDistribution[21], 0.025]
```

```
-2.07961
```

Make sure to recompute the critical score each time, because it depends on your degrees of freedom.

2. find corresponding tail area for desired confidence level



Example: Construct the confidence interval for $\bar{x}$

biscuit intake	$\bar{x}$	s	n
solitaire	52.1 g	45.1 g	22
no distraction	27.1 g	26.4 g	22

solitaire: $\bar{x} \pm t^* \text{ SE} = 52.1 \pm 2.08 \times \frac{45.1}{\sqrt{22}}$

$= 52.1 \pm 2.08 \times 9.61535 = 52.1 \pm 19.9999$

$= 52.1 \pm 20 = (32.1, 72.1)$

“We are 95% confident that distracted eaters consume between 32.1 and 72.1 g of snacks post-meal.

no distraction: $\bar{x} \pm t^* \text{ SE} = 27.1 \pm 2.08 \times \frac{26.4}{\sqrt{22}}$

$= 27.1 \pm 2.08 \times 5.6285 = 27.1 \pm 11.7073$

$= (15.3927, 38.8073)$

"We are 95% confident that undistracted eaters consume between 15.4 and 38.8 g of snacks post-meal.

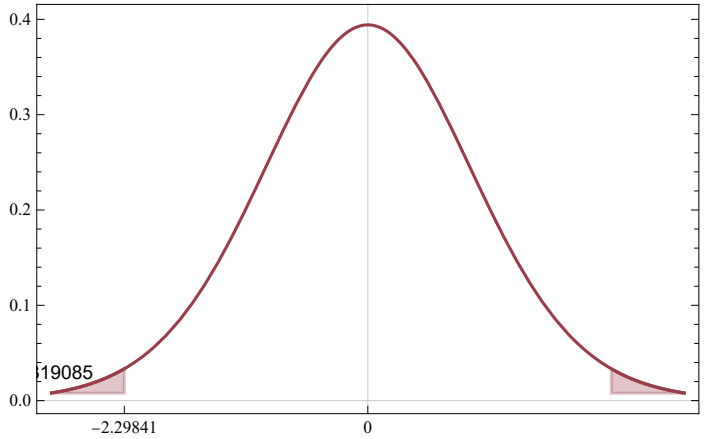
Example

Do these data provide convincing evidence that the amount of snacks consumed by distracted eaters post-lunch is different than the suggested serving size of 30 g?

biscuit intake	$\bar{x}$	s	n
solitaire	52.1 g	45.1 g	22
no distraction	27.1 g	26.4 g	22

Out[•]=

H_0:μ=52.1
H_A:$\mu\neq$52.1
$\bar{x}$=30
SE=9.61535
n=22
$\bar{x}_{\text{diff}} \approx f_{\text{Student}-t}(\text{dof}=21)$



$T=\frac{30-52.1}{\frac{9.61535}{\sqrt{22}}}=-2.29841$
p−value=0.0319085

Compare 95% CI (32.1,72.1)
The data provide convincing evidence to reject H_0.

Conditions?

- Independent?
  - random assignment & n<10% of population OK
  - sample size / distribution skew?

Inference for comparing two small sample means

Confidence interval

Out[ ]//TraditionalForm=

$$(\bar{x}_1 - \bar{x}_2) \pm (t_{\text{df}})^* \, \text{SE}_{(\bar{x}_1 - \bar{x}_2)}$$

Out[ ]//TraditionalForm=

$$\text{SE} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Hypothesis test

Out[ ]//TraditionalForm=

$$T_{\text{df}} = \frac{\text{obs} - \text{null}}{\text{SE}} = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\text{SE}_{(\bar{x}_1 - \bar{x}_2)}}$$

DF for t statistic for inference on difference on two means:

Out[ ]//TraditionalForm=

$$\text{df} = \min(n_1 - 1, n_2 - 1)$$

Example

Do these data provide convincing evidence of a difference between the average post-meal snack consumption between those who eat with and without distractions?

biscuit intake	$\bar{x}$	s	n
solitaire	52.1 g	45.1 g	22
no distraction	27.1 g	26.4 g	22

Out[•]=

$H_0: \mu = 0$

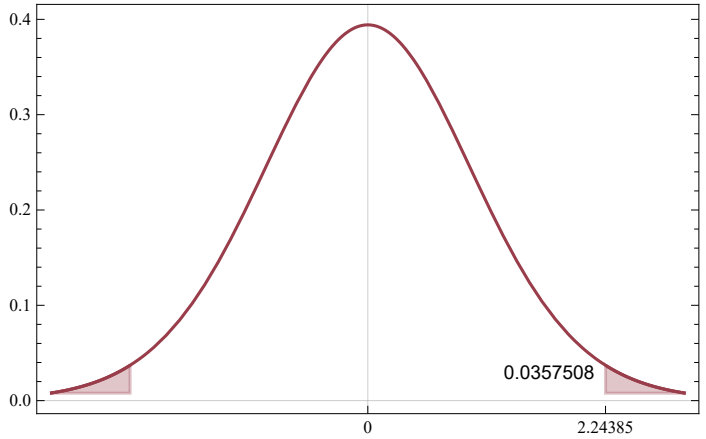
$H_A: \mu \neq 0$

$\bar{x} = 25$

SE = 11.1416

n = 22

$\bar{x}_{diff} \approx f_{Student-t}(dof = 21)$



$T = \frac{25-0}{\frac{11.1416}{\sqrt{22}}} = 2.24385$

$p-value = 0.0357508$

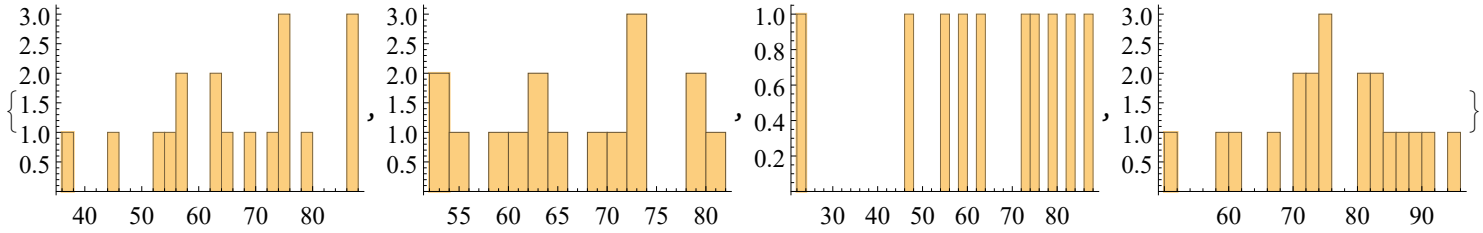
What is the difference between the average post-meal snack consumption between those who eat with and without distractions?

$(\bar{x}_{wd} - \bar{x}_{wod}) \pm t^*$ SE = $25 \pm 2.08 \times 11.14 = 25 \pm 23.17 = (1.83, 48.17)$

Comparing more than two means

Introduction

Example: In the summer term 2022, 64 physics student attend the experimental physics 1 exercises. They were distributed across 4 groups. The Table shows the average group results (max 100 points).



	ÜG1	ÜG2	ÜG3	ÜG4
$\bar{x}$	66.811	66.6052	64.2458	76.3887
n	18	16	10	20
s	14.392	9.32177	19.5919	11.3386

Is there a difference between the average score from different groups?

- To compare means of 2 groups we use a Z or T statistic.
- To compare means of 3+ groups we use a test called analysis of variances (ANOVA) and a new statistic called F.

ANOVA

H_0 : The mean outcome is the same across all categories : $\mu_1 = \mu_2 = \ldots = \mu_k$

H_A : At least one pair of means are different from each other.

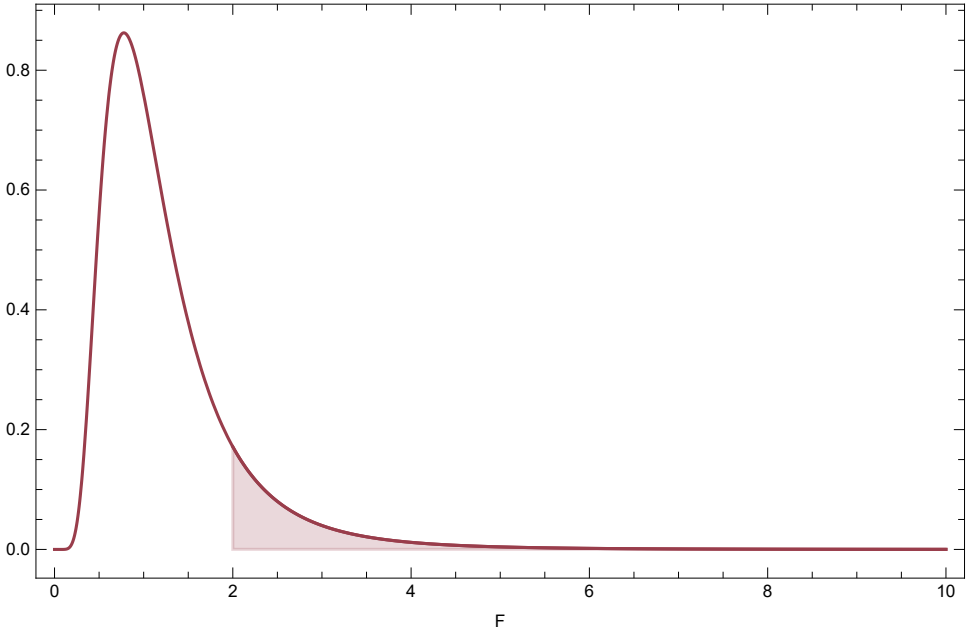
μ_i : mean of the outcome for category i

k : number of groups

test statistics

Out[ ]//TraditionalForm=

$$F = \frac{\text{variability betw. groups}}{\text{variability within groups}}$$



F-(ratio) distribution

AKA: Snedecor's F distribution or the Fisher–Snedecor distribution.

PDF - probability density function

Out[ ]=

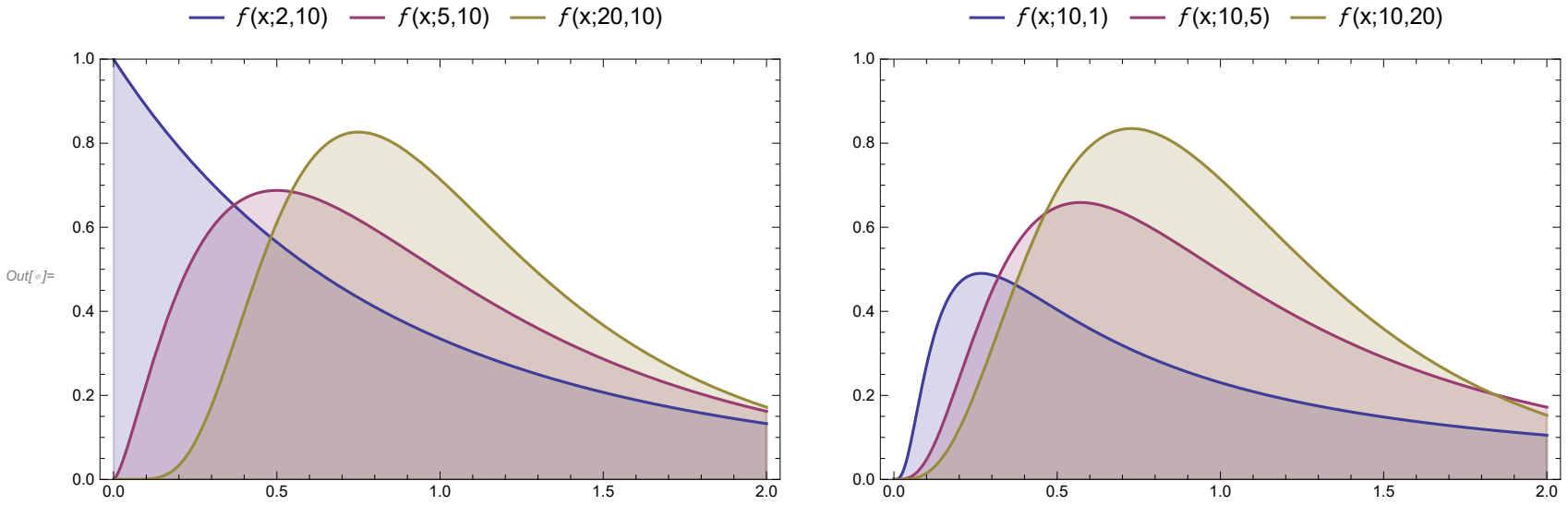
$$f(x\,;d_1,d_2) = \frac{\sqrt{\frac{(d_1 x)^{d_1} d_2^{d_2}}{(d_1 x+d_2)^{d_1+d_2}}}}{x\,B\left(\frac{d_1}{2},\frac{d_2}{2}\right)} = \frac{\left(\frac{d_1}{d_2}\right)^{\frac{d_1}{2}} x^{\frac{d_1}{2}-1}\left(1+\frac{d_1 x}{d_2}\right)^{-\frac{d_1+d_2}{2}}}{B\left(\frac{d_1}{2},\frac{d_2}{2}\right)}$$

and

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

where B is the Beta function and I is the regularized incomplete beta function.

Out[ ]=



A random variate of the F-distribution with parameters d_1 and d_2 arises as the ratio of two appropriately scaled chi-squared variates

Out[ ]= $X = \frac{U_1/d_1}{U_2/d_2}$ where U_1 and U_2 have chi-squared distributions with d_1 and d_2 degrees of freedom

ANOVA

The underlying idea of ANOVA is variability partitioning, i.e. we split the total variability in our data, e.g. the physics score, into:

1. between group variability: the variability attributed to the respective class, e.g. the Übungsgruppe.
2. within group variability attributed to other factors.

ANOVA

		Df	Sum Sq	Mean Sq	F value	Pr(>F)
Group	ÜGs	3	1487.99	495.996	2.77559	0.0489682
Error	Residuals	60	10722.	178.7		
	Total	63	12210.			

ANOVA

		Df	Sum Sq	Mean Sq	F value	Pr(>F)
Group	ÜGs	3	1487.99	495.996	2.77559	0.0489682
Error	Residuals	60	10 722.	178.7		
	Total	63	12 210.			

Sum of squares total (SST)

- measures the **total variability** in the response variable
- calculated similar to variance (without scaling by sample size)

Out[•]=

$$SST = \sum_{1=1}^{n} (y_i - \overline{y})^2$$

y_i: value of response variable for EACH observation
$\overline{y}$: grand mean of response variable

$SST = (78.7 - 69.35)^2 + (75.7 - 69.35)^2 + ...+(83.4 - 69.35)^2 \approx 12\,210$

ANOVA

		Df	Sum Sq	Mean Sq	F value	$\Pr(>F)$
Group	ÜGs	3	1487.99	495.996	2.77559	0.0489682
Error	Residuals	60	10722.	178.7		
	Total	63	12210.			

Sum of squares group (SSG)

- measures the variability **between groups**
- **explained variability:** deviation of group mean from overall mean, weighted by sample size

Out[•]=

$$SSG = \sum_{j=1}^{k} n_j \left(\bar{y}_j - \bar{y}\right)^2$$

n_j: number of observations in group j

$\bar{y}_j$: mean of the response variable for group j

$\bar{y}$: grand mean of response variable

$$SSG = 18\,(66.811 - 69.35)^2 + 16\,(66.6052 - 69.35)^2 + 10\,(64.2458 - 69.35)^2 + 20\,(76.3887 - 69.35)^2 = 1487.99$$

ANOVA

		Df	Sum Sq	Mean Sq	F value	Pr(>F)
Group	ÜGs	3	1487.99	495.996	2.77559	0.0489682
Error	Residuals	60	10722.	178.7		
	Total	63	12210.			

Sum of squares error (SSE)

- measures the variability **within groups**
- **unexplained variability:** unexplained by the group variable, due to other reasons

Out[ ]//TraditionalForm=

$$SSE = SST - SSG$$

SSE = 12210 – 1487.99 = 10722.

ANOVA

		Df	Sum Sq	Mean Sq	F value	Pr(>F)
Group	ÜGs	3	1487.99	495.996	2.77559	0.0489682
Error	Residuals	60	10722.	178.7		
	Total	63	12210.			

To get a measure of the average variabilities (not the total) we need to scale by a measure that incorporates sample sizes and number of group ⟶ **degrees of freedom**

ANOVA

		Df	Sum Sq	Mean Sq	F value	Pr(>F)
Group	ÜGs	3	1487.99	495.996	2.77559	0.0489682
Error	Residuals	60	10722.	178.7		
	Total	63	12210.			

degrees of freedom associated with ANOVA

- total: $df_T = n - 1$ ⟶ example: 64 - 1 = 63
- group: $df_G = k - 1$ ⟶ example: 4 - 1 = 3
- error: $df_E = df_T - df_G$ ⟶ example: 63 - 3 = 60

ANOVA

		Df	Sum Sq	Mean Sq	F value	Pr(>F)
Group	ÜGs	3	1487.99	495.996	2.77559	0.0489682
Error	Residuals	60	10722.	178.7		
	Total	63	12210.			

Mean square error

Average variability **between** and **within** groups, calculated as the total variability (sum of squares) scaled by the associated degrees of freedom.

- group: $\text{MSG} = \text{SSG} / \text{df}_G \longrightarrow$ example: 1487.99 / 3 = 495.996
- error: $\text{MSE} = \text{SSE} / \text{df}_E \longrightarrow$ example: 10722. / 60 = 178.7

ANOVA

		Df	Sum Sq	Mean Sq	F value	Pr(>F)
Group	ÜGs	3	1487.99	495.996	2.77559	0.0489682
Error	Residuals	60	10722.	178.7		
	Total	63	12210.			

F Statistic

Ratio of the **between group** and **within group** variability:

Out[•]//TraditionalForm=

$$F = \frac{\text{MSG}}{\text{MSE}}$$

example: 495.996/178.7=2.7755791829882486

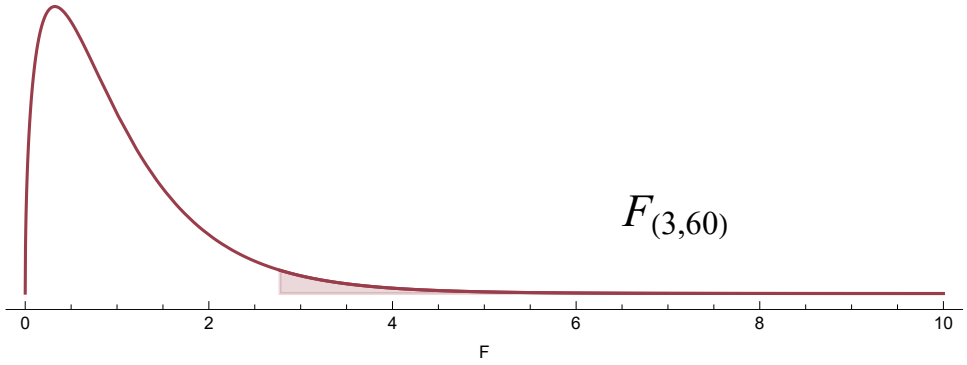
ANOVA

		Df	Sum Sq	Mean Sq	F value	$\mathbb{P}$(>F)
Group	ÜGs	3	1487.99	495.996	2.77559	0.0489682
Error	Residuals	60	10722.	178.7		
	Total	63	12210.			

p-Value

- **p-value** is the probability of at least as large a ratio between the "between" and "within" group variabilities if in fact the means of all groups are equal

- area under the F curve, with degrees of freedom df_G and df_E, above the observed F statistic



use a web applet to compute p: http://bitly.com/dist_calc or R: pf(f, dfg, dfe, lower.tail = FALSE), or *Mathematica*: 1-CDF[FRatioDistribution[dfg,dfe],f]

ANOVA

		Df	Sum Sq	Mean Sq	F value	$\Pr(>F)$
Group	ÜGs	3	1487.99	495.996	2.77559	0.0489682
Error	Residuals	60	10722.	178.7		
	Total	63	12210.			

Conclusion

- If p < α then reject H_0: The data provide convincing evidence that at least one pair of population means are different from each other (but we can't tell which one).
- If p≥ α then fail to reject H_0 : The data do not provide convincing evidence that one pair of population means are different from each other, the observed differences in sample means are attributable to sampling variability (or chance).

ANOVA

- The advantage of the ANOVA F-test is that we do not need to pre-specify which groups are to be compared, and we do not need to adjust for making multiple comparisons.
- The disadvantage of the ANOVA F-test is that if we reject the null hypothesis, we do not know which groups can be said to be significantly different from the others, nor, if the F-test is performed at level α we can state that the group pair with the greatest mean difference is significantly different at level α.

Assumptions:

- **Independence of observations** – this is an assumption of the model that simplifies the statistical analysis.
  - within groups: sampled observations must be independent
  - between groups: the groups must be independent of each other (non–paired)
- **Normality** – the distributions of the residuals are normal.
- Equality (or "homogeneity") of variances, called **homoskedasticity** — the variance of data in groups should be the same (i.e roughly equal).

Init