Boring but important disclaimers:

- If you are not getting this from the GitHub repository or the associated Canvas page (e.g. CourseHero, Chegg etc.), you are probably getting the substandard version of these slides Don't pay money for those, because you can get the most updated version for free at

  `https://github.com/julianmak/academic-notes`

  The repository principally contains the compiled products rather than the source for size reasons.
- Associated Python code (as Jupyter notebooks mostly) will be held on the same repository. The source data however might be big, so I am going to be naughty and possibly just refer you to where you might get the data if that is the case (e.g. JRA-55 data). I know I should make properly reproducible binders etc., but I didn't...
- I do not claim the compiled products and/or code are completely mistake free (e.g. I know I don't write Pythonic code). Use the material however you like, but use it at your own risk.
- As said on the repository, I have tried to honestly use content that is self made, open source or explicitly open for fair use, and citations should be there. If however you are the copyright holder and you want the material taken down, please flag up the issue accordingly and I will happily try and swap out the relevant material.

OCES 3301 :

basic **Data Analysis** in ocean sciences

Session 5: statistical tests

Outline

(Just overview here; for actual content see Jupyter notebooks)

- brief introduction to probability

  → pdfs, Gaussian, law of large numbers, CLT

  → Confidence Intervals

- hypothesis testing (mostly focus on *analysis*)

  → null hypothesis

  → confidence threshold

  → computing the test statistic (Z-test example), p-values

  → banana skins

  → Type I and II errors, statistical power (cf. experimental design)

DISCLAIMER

- so I hate most of this stuff and I almost never use it

  $\rightarrow$ I don’t use statistical testing in my work

DISCLAIMER

- so I hate most of this stuff and I almost never use it
  $\rightarrow$ I don't use statistical testing in my work

- this stuff is / can be very confusing

DISCLAIMER

- so I hate most of this stuff and I almost never use it
  $\rightarrow$ I don't use statistical testing in my work

- this stuff is / can be very confusing
  $\rightarrow$ you are not alone...

DISCLAIMER

- so I hate most of this stuff and I almost never use it
  $\rightarrow$ I don't use statistical testing in my work

- this stuff is / can be very confusing
  $\rightarrow$ you are not alone...
  $\rightarrow$ be aware of the (many) banana skins

DISCLAIMER

- so I hate most of this stuff and I almost never use it

  $\rightarrow$ I don't use statistical testing in my work

- this stuff is / can be very confusing

  $\rightarrow$ you are not alone...

  $\rightarrow$ be aware of the (many) banana skins

- I think there are better tools to use (Bayesian formulations, confidence intervals), but not touching those here

  $\rightarrow$ the statistical tests here are more **classical**

  $\rightarrow$ arguably more technical (not really though...)

DISCLAIMER

- so I hate most of this stuff and I almost never use it
  $\rightarrow$ I don't use statistical testing in my work

- this stuff is / can be very confusing
  $\rightarrow$ you are not alone...
  $\rightarrow$ be aware of the (many) banana skins

- I think there are better tools to use (Bayesian formulations, confidence intervals), but not touching those here
  $\rightarrow$ the statistical tests here are more **classical**
  $\rightarrow$ arguably more technical (not really though...)

- not going to talk about experimental design, though it is probably more important than the statistical analysis

Lasciate ogne speranza, voi ch'intrate

Figure: Cursed image.

Motivation: sea cucumber

Figure: Moldy sea cucumber.

Suppose sea cucumber has some distribution of weight that we can measure:

Q. does change in diet affect her weight?

Q. does exercise regime affect her weight?

- might expect to, but how do we distinguish noise (e.g. natural random fluctuations) with “real” effect?

Motivation: sea cucumber

Figure: Moldy sea cucumber.

e.g. say from samples,

$$\{\mu_1 = 3.00, \quad \sigma_1 = 0.5\}$$

$$\{\mu_2 = 3.20, \quad \sigma_2 = 0.5\}$$

- mean is different, so has effect?
  $\rightarrow$ but could just be a fluke?

- hypothesis testing as a tool to say whether differences are statistically significant

Some probability

- assign some real value between 0 and 1 to some event
  $\rightarrow$ 0 is never, 1 is certainly
- e.g. throwing a fair coin, two events, expect $p = 1/2$

Q. that's in principle, what if you actually try it?

Some probability

- assign some real value between 0 and 1 to some event
  $\rightarrow$ 0 is never, 1 is certainly
- e.g. throwing a fair coin, two events, expect $p = 1/2$

Q. that's in principle, what if you actually try it?

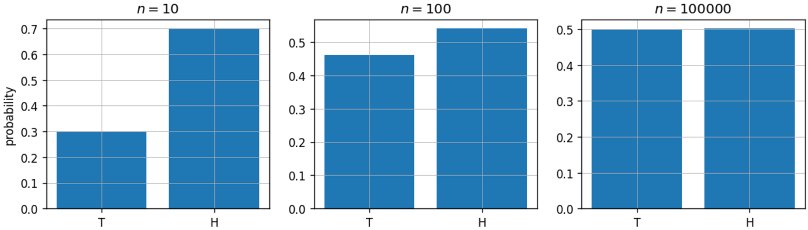


Figure: Result of hypothetical coin tosses displayed in bar graph.

Some probability

- going back to the sea cucumber, we might have a sample of weights and display it in a histogram

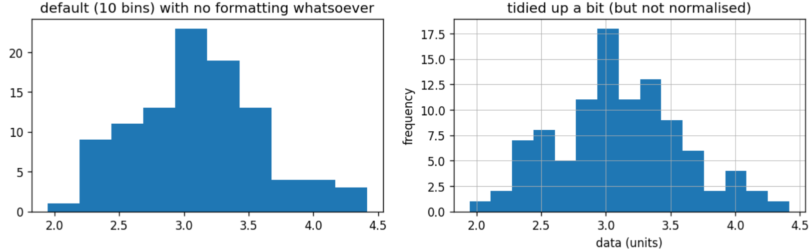


Figure: Result of hypothetical sea cucumber weight.

- obtained form binning procedure
  $\rightarrow$ probability is related here to **integral** of the graph

Some probability

- histogram related to the probability distribution function (**pdf**) of the underlying sample

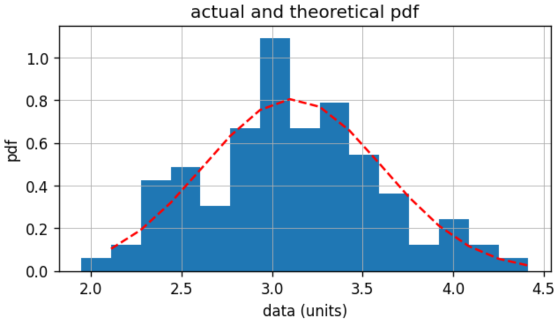


Figure: Result of hypothetical sea cucumber weight with Gaussian pdf.

Q. where did I get the red line from though?

Some probability

Gaussian or normal distribution has pdf

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]$$

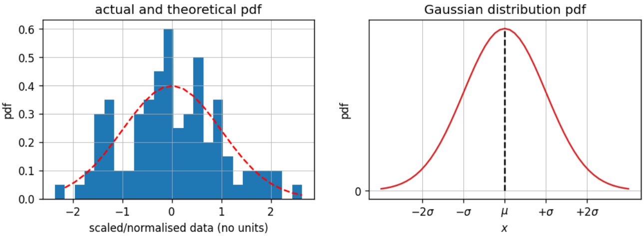


Figure: The Gaussian pdf (with units deliberately omitted). Obtain probability from an integral.

- approximate population (μ, σ) from sample $(\bar{x}, s)$

Q. what would the pdf of the uniform distribution look like?

Some probability

Point here is that if you have the pdf you have basically everything

Some probability

Point here is that if you have the pdf you have basically everything

Theorem (Central Limit Theorem (CLT))

(Very loosely) If your sample is large enough, under fairly general conditions (!) you can approximate most data distributions as a Gaussian distribution, even if the underlying distribution is not necessarily Gaussian.

Some probability

Point here is that if you have the pdf you have basically everything

Theorem (Central Limit Theorem (CLT))

(Very loosely) If your sample is large enough, under fairly general conditions (!) you can approximate most data distributions as a Gaussian distribution, even if the underlying distribution is not necessarily Gaussian.

- for large enough samples, you can fit it to a Gaussian pdf...

Some probability

Point here is that if you have the pdf you have basically everything

Theorem (Central Limit Theorem (CLT))

(Very loosely) If your sample is large enough, under fairly general conditions (!) you can approximate most data distributions as a Gaussian distribution, even if the underlying distribution is not necessarily Gaussian.

- for large enough samples, you can fit it to a Gaussian pdf...
- ...and if you have the pdf you have basically everything!

Some probability

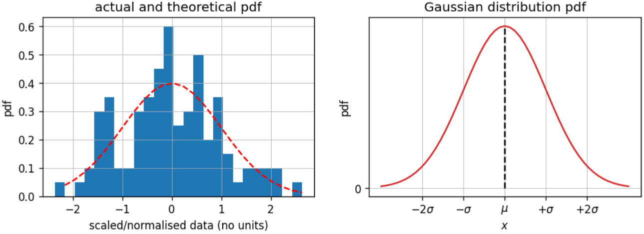


Figure: The Gaussian pdf (with units deliberately omitted). Obtain probability from an integral.

- **68-95-99.7 rule**, 68, 95 and 99.7% of the data lies within 1, 2 and 3 s.t.d. of the mean

  $\rightarrow$ whenever CLT applies (which is quite often!)

Some probability

e.g.

$$p(-\sigma < z < \sigma) = \frac{1}{\sqrt{2\pi}} \int_{-\sigma}^{\sigma} \mathrm{e}^{-z^2/2}\, \mathrm{d}z \approx 0.68$$

Some probability

e.g.

$$p(-\sigma < z < \sigma) = \frac{1}{\sqrt{2\pi}} \int_{-\sigma}^{\sigma} \mathrm{e}^{-z^2/2}\, \mathrm{d}z \approx 0.68$$

Flip it around: for the some given P, find the $\tilde{z}$ such that

$$p(-\tilde{z} < z < \tilde{z}) = \frac{1}{\sqrt{2\pi}} \int_{-\tilde{z}}^{\tilde{z}} \mathrm{e}^{-z^2/2}\, \mathrm{d}z = P.$$

Some probability

e.g.

$$p(-\sigma < z < \sigma) = \frac{1}{\sqrt{2\pi}} \int_{-\sigma}^{\sigma} \mathrm{e}^{-z^2/2}\, \mathrm{d}z \approx 0.68$$

Flip it around: for the some given P, find the $\tilde{z}$ such that

$$p(-\tilde{z} < z < \tilde{z}) = \frac{1}{\sqrt{2\pi}} \int_{-\tilde{z}}^{\tilde{z}} \mathrm{e}^{-z^2/2}\, \mathrm{d}z = P.$$

Confidence Interval

- the interval that contains P amount of probability

  $\rightarrow$ e.g. **95% confidence interval** for Gaussian data would be around $(-2\sigma, 2\sigma)$ $(-1.96\sigma, 1.96\sigma)$ is more accurate but whatever...

Some probability

Z-score or standardised scores: with samples x_i, define

$$z_i = \frac{x_i - \mu}{\sigma}.$$

- essentially re-scaled Gaussian
  $\rightarrow$ cf. what was done for the PCA two sessions ago
  $\rightarrow$ allows somewhat of a like-for-like comparison

Back to the sea cucumber

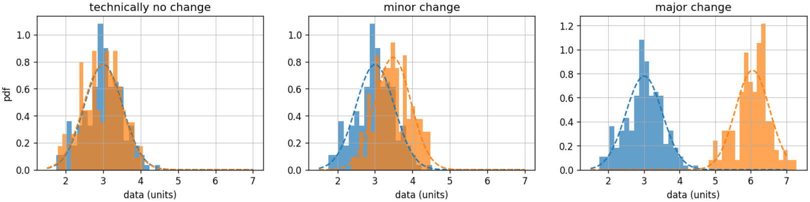


Figure: Control and varied sample distributions and associated Gaussian pdf.

- we are basically dealing with samples (going to assume CLT holds here)

Q. variability in data always exist, so how to distinguish change from noise?

Back to the sea cucumber

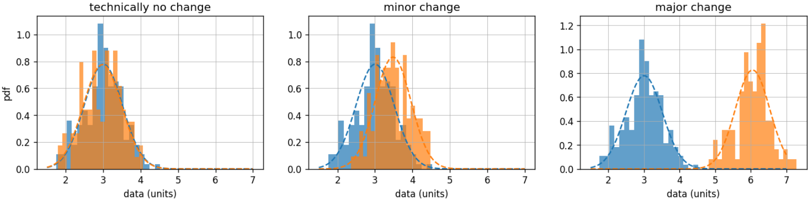


Figure: Control and varied sample distributions and associated Gaussian pdf.

- ▶ we are basically dealing with samples (going to assume CLT holds here)

Q. variability in data always exist, so how to distinguish change from noise?

? non-overlapping confidence intervals?

$\rightarrow$ quite strict (and **under-powered**; see later)

Hypothesis testing

A more standard and routine (doesn't mean it's a good thing necessarily):
hypothesis testing and computing test statistics

Hypothesis testing

A more standard and routine (doesn't mean it's a good thing necessarily): hypothesis testing and computing test statistics

cf. proof by contradiction: want to proof X, so

- assume *not* X
- start from not X, logically derive consequences, avoiding illegal logical manoeuvres
- come to a contradiction
- if no illegal manoeuvres, then initial assumption must be false, and there X is true ∎

Hypothesis testing

Hypothesis testing is similar (though logically weaker in some sense):

Start with a null hypothesis H_0 (opposite to what you want to show usually)

- assume H_0

Hypothesis testing

Hypothesis testing is similar (though logically weaker in some sense):

Start with a null hypothesis H_0 (opposite to what you want to show usually)

- assume H_0
- decide test and significance level (depends on the thing you want to show)

Hypothesis testing

Hypothesis testing is similar (though logically weaker in some sense):

Start with a null hypothesis H_0 (opposite to what you want to show usually)

- assume H_0
- decide test and significance level (depends on the thing you want to show)
- compute test statistics (depends on test)

Hypothesis testing

Hypothesis testing is similar (though logically weaker in some sense):

Start with a null hypothesis H_0 (opposite to what you want to show usually)

- assume H_0
- decide test and significance level (depends on the thing you want to show)
- compute test statistics (depends on test)
- if associated probability of computed test statistic is low, then it is either:
  1. a really surprising result

Hypothesis testing

Hypothesis testing is similar (though logically weaker in some sense):

Start with a null hypothesis H_0 (opposite to what you want to show usually)

- assume H_0
- decide test and significance level (depends on the thing you want to show)
- compute test statistics (depends on test)
- if associated probability of computed test statistic is low, then it is either:
    1. a really surprising result
    2. or H_0 is incompatible with data

Hypothesis testing

Hypothesis testing is similar (though logically weaker in some sense):

Start with a null hypothesis H_0 (opposite to what you want to show usually)

- assume H_0
- decide test and significance level (depends on the thing you want to show)
- compute test statistics (depends on test)
- if associated probability of computed test statistic is low, then it is either:
  1. a really surprising result
  2. or H_0 is incompatible with data
- if latter, reject H_0, and there is statistical evidence in support for *not* H_0 (which is the thing you wanted anyway)

Hypothesis testing

e.g. sea cucumber, want to know if diet has any effect on weight

Hypothesis testing

e.g. sea cucumber, want to know if diet has any effect on weight

H_0 : diet has **NO** bearing on weight

Hypothesis testing

e.g. sea cucumber, want to know if diet has any effect on weight

H_0 : diet has **NO** bearing on weight

test : large enough samples, assume Gaussian statistics, do two-tailed **Z-test**

Hypothesis testing

e.g. sea cucumber, want to know if diet has any effect on weight

H_0 : diet has **NO** bearing on weight

test : large enough samples, assume Gaussian statistics, do two-tailed **Z-test**

α : choose $\alpha = 0.05$

$\rightarrow$ how far you are in tails of the pdf

$\rightarrow$ for Gaussian pdf, corresponds to Z-score of around 2, because 95% CI is around $(-2\sigma, +2\sigma)$

Hypothesis testing

e.g. sea cucumber, want to know if diet has any effect on weight

H_0 : diet has **NO** bearing on weight

test : large enough samples, assume Gaussian statistics, do two-tailed **Z-test**

α : choose $\alpha = 0.05$

$\rightarrow$ how far you are in tails of the pdf

$\rightarrow$ for Gaussian pdf, corresponds to Z-score of around 2, because 95% CI is around $(-2\sigma, +2\sigma)$

compute : compute Z-statistics (see notebook)

Hypothesis testing

e.g. sea cucumber, want to know if diet has any effect on weight

H_0 : diet has **NO** bearing on weight

test : large enough samples, assume Gaussian statistics, do two-tailed **Z-test**

α : choose $\alpha = 0.05$

$\rightarrow$ how far you are in tails of the pdf

$\rightarrow$ for Gaussian pdf, corresponds to Z-score of around 2, because 95% CI is around $(-2\sigma, +2\sigma)$

compute : compute Z-statistics (see notebook)

conclude : if Z-statistic large or corresponding p-value small, then reject H_0

(see actual code syntax in notebook)

Z-test (demonstration of sorts here)

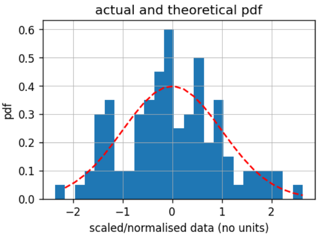
- Z-test calculates Z-statistic with sample mean $\overline{x}$

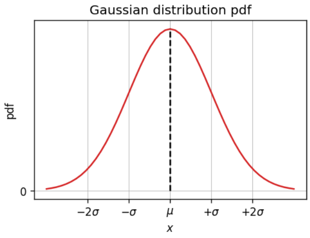
$$Z = \frac{\overline{x} - \mu}{\sigma / \sqrt{N}}$$

$\rightarrow$ with H_0 the mean is the **same**

$\rightarrow$ if Z big enough (so the associated p-value is small) then evidence to reject H_0

$\rightarrow$ Gaussian, large samples, known σ (could approximate with s)





Easy right?

Easy right?

Figure: Banana says no.

- banana skins in no particular order (there are so many of them)...

Banana skin 1: H_0

What if you fail to reject the null hypothesis?

Banana skin 1: H_0

What if you fail to reject the null hypothesis?

- that's it, **you fail to reject the null hypothesis**, no more and no less

Banana skin 1: H_0

What if you fail to reject the null hypothesis?

- that's it, **you fail to reject the null hypothesis**, no more and no less

  $\rightarrow$ this does NOT mean H_0 is true, nor that being able to reject H_0 means H_0 is false

  $\rightarrow$ it really just means you can't say anything

Banana skin 1: H_0

What if you fail to reject the null hypothesis?

- that's it, **you fail to reject the null hypothesis**, no more and no less

  $\rightarrow$ this does NOT mean H_0 is true, nor that being able to reject H_0 means H_0 is false

  $\rightarrow$ it really just means you can't say anything

- cf. proof by contradiction: not being able to find a contradiction could mean

  $\rightarrow$ there really is nothing there

  $\rightarrow$ you aren't looking hard enough

Banana skin 2: p-values and H_0

For $\alpha = 0.05$ and I reject H_0, so test tells me H_0 is only 5% likely to be true

Banana skin 2: p-values and H_0

For $\alpha = 0.05$ and I reject H_0, so test tells me H_0 is only 5% likely to be true

- (frequentist point of view) H_0 is either true or false, it can't be 5% true
- probability of $p = 0.05$ is
  $\times$ hypothesis given data $p(H_0|x_i)$
  ✓ data given hypothesis $p(x_i|H_0)$
- rule of thumb (frequentist view): don't assign probabilities to hypotheses

Banana skin 3: p-values

$\alpha = 0.05$ **and I reject** H_0**, so what I observed would have a probability of** 5% **that is was due to noise**

Banana skin 3: p-values

$\alpha = 0.05$ **and I reject** H_0**, so what I observed would have a probability of** 5% **that is was due to noise**

- again no, since p-value is tagged with the null hypothesis, i.e., I would have observed this signal *given* the null hypothesis

Banana skin 4: *p*-values

$\alpha = 0.05$ **is the gold standard**

Banana skin 4: p-values

$\alpha = 0.05$ **is the gold standard**

- $10^{10^{10}}$ NO!!!

Banana skin 4: p-values

$\alpha = 0.05$ **is the gold standard**

- $10^{10^{10}}$ NO!!!

  $\rightarrow$ comes from (Ronald) Fisher's paper in the 30s or so, when dealing with small samples (≈ 20?)

  $\rightarrow$ can easily be abused to generate false-positives with multiple testing or large sampling (see notebook)

- just a convention (and prone to abuse)

  $\rightarrow$ e.g. particle physics uses 5σ ($\alpha = 0.0000003$, or 1 in 3.5 million, partly because of multiple sampling going on)

Banana skin 5: p-values

My test statistic is large or my p-value is small, so my result is extremely important

Banana skin 5: p-values

My test statistic is large or my p-value is small, so my result is extremely important

- also no

Banana skin 5: p-values

My test statistic is large or my *p*-value is small, so my result is extremely important

- also no

  → statistical significance is not practical significance, p-values can't tell you the latter

  → H_0 statements are very broad: H_0 is no change, not H_0 is there *is* change, but it doesn't tell you how much change

- need interpretation: 1 kg difference in a sea cucumber is not the same as 1 kg difference in a whale

Type I and II errors

Type I errors (false-positives)

- rejecting H_0 when H_0 is true

  $\rightarrow$ related to choice of significance α

Type II errors (false-negatives)

- fail to reject H_0 when H_0 is false

	H_0 true	H_0 false
reject H_0	Type I	✓
fail to reject H_0	✓	Type II

Type I and II errors

Type I errors (false-positives)

- rejecting H_0 when H_0 is true

  $\rightarrow$ related to choice of significance α

Type II errors (false-negatives)

- fail to reject H_0 when H_0 is false

	H_0 true	H_0 false
reject H_0	Type I	✓
fail to reject H_0	✓	Type II

or, with H_0 = someone is innocent,

	innocent	murderer
found guilty	wrongful conviction	✓
found not guilty	✓	fail to prosecute

Type I and II errors

Type II errors β related to statistical power $1 - \beta$

- really to do with experimental design
  $\rightarrow$ choice of sample size to detect effect (see notebook)

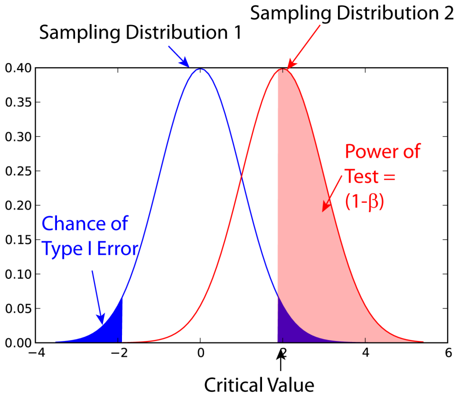


Figure: Graphical demonstration of Type I and II errors. From Wikipedia.

Bad practices 1: “torturing the data until it confesses”

Bad practices 1: “torturing the data until it confesses”

This statistical test didn’t work so I will use another one, or

These are outliers, so I will get rid of those, or

I will sub-sample dataset and try again etc.

Bad practices 1: "torturing the data until it confesses"

This statistical test didn't work so I will use another one, or

These are outliers, so I will get rid of those, or

I will sub-sample dataset and try again etc.

- $\alpha = 0.05$ so 5% of the time you will reject the null hypothesis when you shouldn't have (Type I error, false-positives)

  $\rightarrow$ multiple testing you could be sampling that 5%

  $\rightarrow$ issue of stop when you get a hit (DON'T!!!)

  (see notebook for an example of this)

- leads to false discoveries (see notebook for an example to do with academic publishing)

Bad practices 2: report only reject or not reject

Reporting reject/fail to reject only

- report the full p-value

  $\rightarrow$ above/below boundary is not saying whether hypothesis is true/false (banana skin 2)

  $\rightarrow$ the threshold is a convention (banana skin 4)

Bad practices 3: being hung up on the analysis

The analysis part is so hard I need to pay most of my attention to it

- if your experimental design is faulty than one would hope (!) that no amount of wizardry will fix that...

  $\rightarrow$ have a think about the validity of design and tools also

Jupyter notebook

Probably more but this is surely getting tedious... go to 05 Jupyter notebook to get some code practise

- RNGesus and probability
- written overview of hypothesis testing

  $\rightarrow$ one example using Z-tests

Statistics is just a <u>tool</u>, no more and no less

- YOU are the user and the onus is on YOU to know enough about to tool to not abuse it