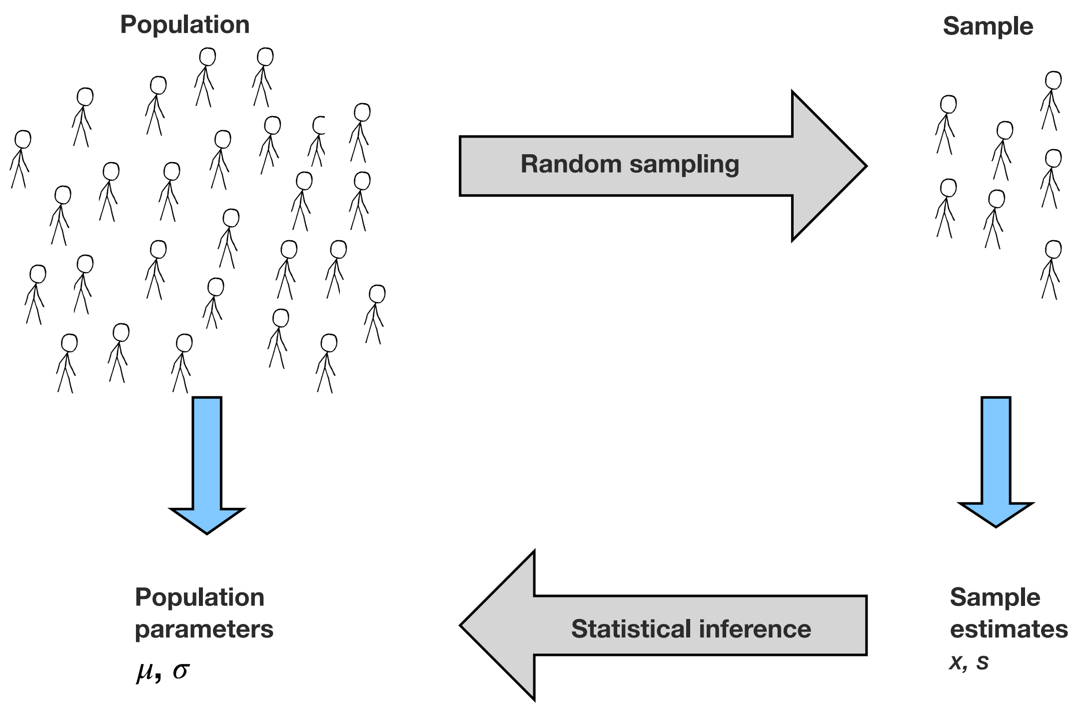
Introduction to hypothesis testing

BIOL01301 Spring 2020

Dr. Spielman

Statistical inference: we don't know the population



Hypothesis testing

- Compare data to the expectation of a **specific null hypothesis**
- Also known as "NHST" = null hypothesis significance testing
- Tests ask: What is the probability of observing *my data or more extreme** under the assumption that the null is TRUE?
  - *This probability = P-value*

The test we use depends on our *question* and *type of data*

- *t*-test compares **means of numeric data**
  - One-sample *t*-test: *My sample mean equals NULL VALUE.*
  - Two-sample *t*-test: *My two samples have equal means.*

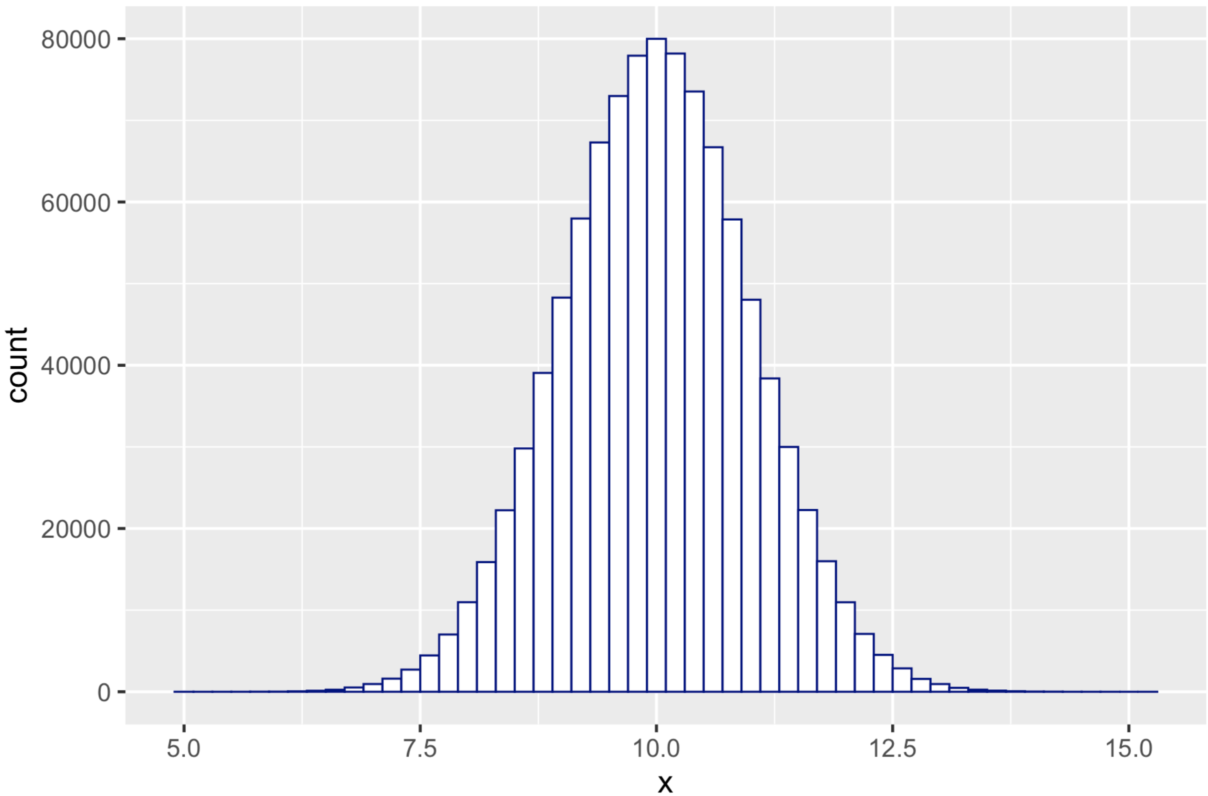
- Chi-squared goodness of fit compares **proportions** (count ratios)
  - *Ratio of counts for these three categories is X:Y:Z*
  - Also, Fisher's Exact Test

- Contingency table analyses look for <u>categorical</u> variable associations
  - *Variable 1 is not associated with variable 2*

All tests employ a *test statistic distribution* to model the <u>null</u>

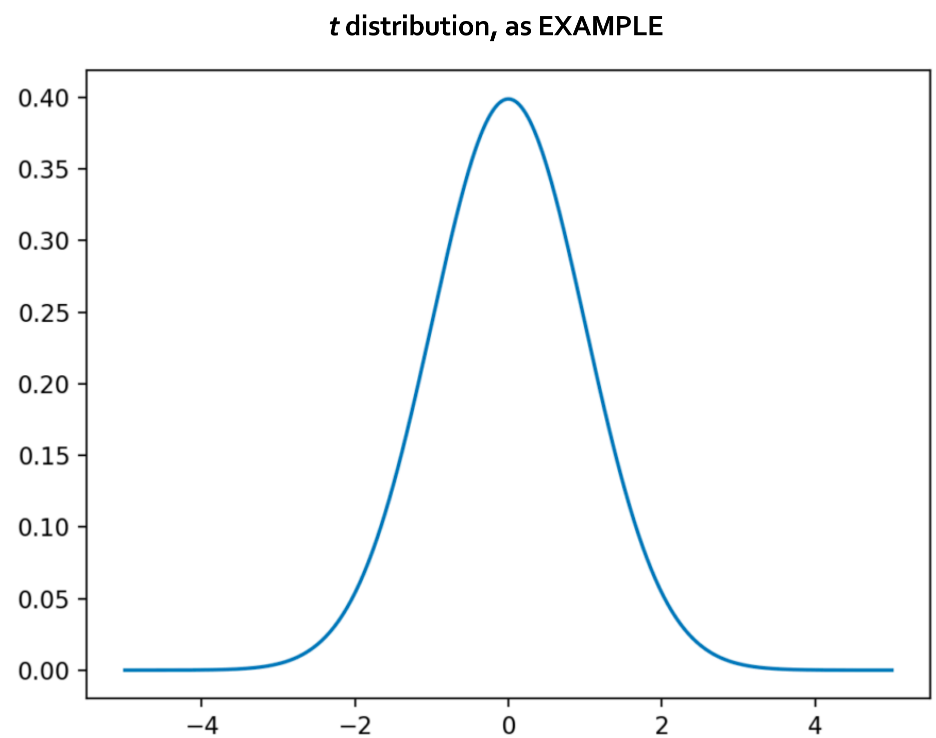
- **Statistic = a value you can calculate from your data (yes, really)**
- *t*-tests consider the *t* statistic, chi-squared tests use the *Chi-squared* statistic
- [technically, the null = sampling distribution of the null hypothesis]

Probability-thinking with distributions



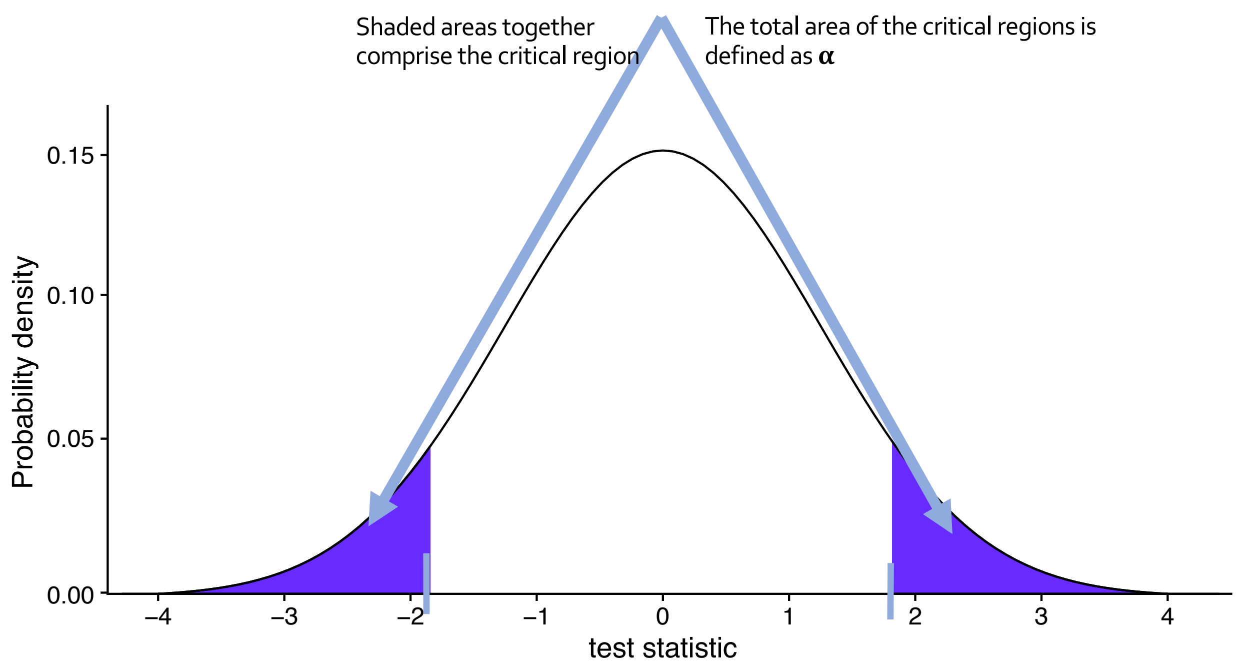
How to do a hypothesis test, *very* generally speaking

- Think: What is your question? Therefore, what test to use?
- *Pick your level of significance aka critical value*
- From your data, calculate the test statistic using the relevant formula (computer!)
- Get the area *towards the tail(s)* of your null
  - One-sided or two-sided?



The P-value is the area under the curve for your test statistic

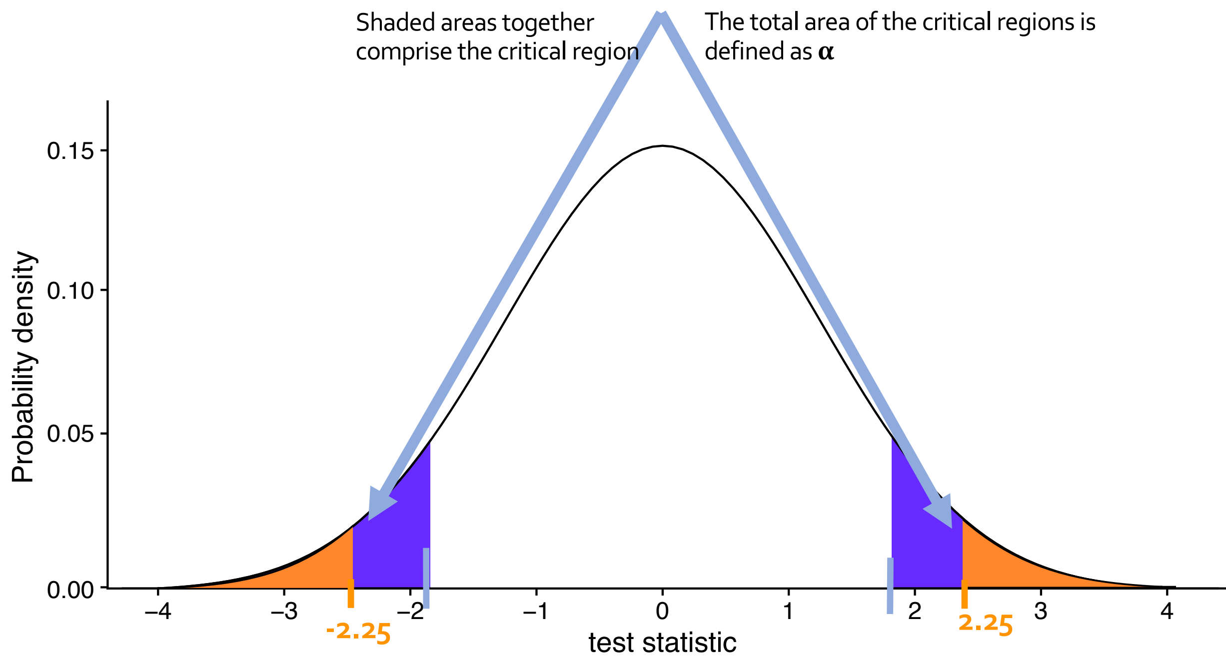
- Result of hypothesis test is **significant** if test statistic falls in the critical region



NOTE: It will not always be symmetric

The P-value is the area under the curve for your test statistic

- Result of hypothesis test is **significant** if test statistic falls in the critical region



NOTE: It will not always be symmetric

For **test statistic = 2.25**, the sum of area is less than **α**

Forming conclusions

Based on your *pre-chosen* α:

1. P-value <= α
   - Significant results allow us to reject the null hypothesis and conclude evidence in favor of the alternative hypothesis

2. P-value > α
   - We do not have significant results. We fail to reject the null hypothesis, and we have no evidence in favor of the alternative hypothesis.

The choice of α is totally arbitrary, but usually you will see 0.05 or 0.01

Statistical significance is not the same as effect size (“biological significance ”)

Exemplified with a two-sample t-test!

Error rates

α sets the overall **false positive rate** for our test procedure. If the null is true, we falsely reject the null 5% of the time for $\alpha=0.05$

		Truth about population (generally unknown)	
		Null is true	**Alternative is true**
Conclusion	**Reject null (P<= α)**	False positive (Type I error)	True positive
	Fail to reject null (P> α)	True negative	False negative (Type II Error)

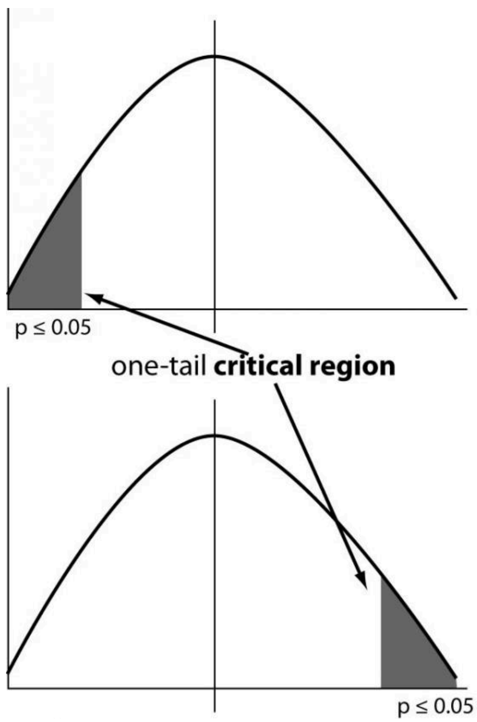
What type of error is it (or is it?)

- A new arthritis drug does not have an effect in clinical trials, even though it actually does reduce arthritis pain.
- A person with HIV receives a positive test result for HIV.
- A person using illegal performing enhancing drugs passes a test clearing them of drug use.
- A study found a significant relationship between neck strain and jogging, when reality there is no relationship.
- A healthy individual gets a positive cancer biopsy result.

One-sided vs. Two-sided (or –tailed)

One-sided tests are **directional**

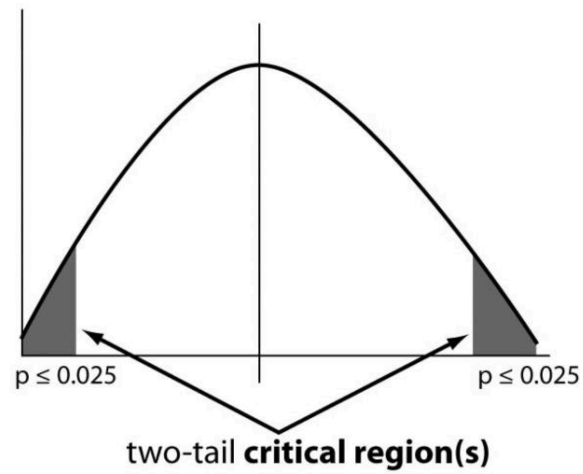
Are my data larger/smaller than null?



<u>TOTAL</u> area = α
no matter what

Two-sided tests are **non-directional**

Do my data <u>differ</u> from null?



What is a P-value?

- The P-value is an area under the curve of the **null distribution**
  - It is therefore the **probability** of observing *this effect or larger* **assuming the null hypothesis is true**

  - **P-value = P(effect or more observed | null is true)**

  - **P-value = 0.009:** If null is true, I would obtain this effect or larger ($t>=2.66$) in 0.9% of such studies due to random sampling error

- A low P-value leads to mental gymnastics that maybe the null is a poor way to think about the situation – something else is going on!
  - We **can never** rule out the possibility that results were fully consistent with null, just unlikely

P-values are not magic

- P-values **cannot** evaluate whether the null or alternative is true
  - Large P-values **do not** prove the null is true
  - Small P-values **do not** prove the alternative is true. They merely ***suggest*** the null perhaps can't explain what we observe.
  - Remember: P-values exist *under assumption that null is true!!!*

- P-values **do not** give the probability that you made the right conclusion

- Two studies with the same P-value do not provide the same weight of evidence

P-values are strongly influenced by sample size

$$t = \frac{\bar{x}-\mu}{\frac{s}{\sqrt{n}}} = \frac{99.59 - 98.6}{\frac{1.44}{\sqrt{\mathbf{15}}}} = 2.66 \rightarrow \text{P=0.009}$$

$$t = \frac{\bar{x}-\mu}{\frac{s}{\sqrt{n}}} = \frac{99.59 - 98.6}{\frac{1.44}{\sqrt{\mathbf{100}}}} = 6.88 \rightarrow \text{P= 2.81e-10}$$

Increasing sample size increases **power**

Power is the probability you detect a **true effect**, i.e. true positive rate

P-values are kind of an accident

Personally, the writer prefers to set a low standard of significance at the 5 percent point . . . . A scientific fact should be regarded as ***experimentally established only if a properly designed experiment rarely fails to give this level of significance.***

- R.A. Fisher

Approach to hypothesis testing

1. Decide what question you are interested in answering
2. Determine the appropriate hypothesis test to use
3. Check that your data meet the assumptions* of the test
4. Compute the *test statistic* for your hypothesis test and the corresponding P-value
5. Draw conclusions using an *a priori* specified α (P-value threshold)

*Parametric only

Assumptions for parametric hypothesis tests

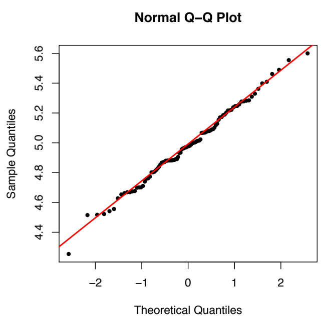
- Data is a *random sample* from the population
- Data is "iid" (independently and identically distributed)
  - All data points are independent from one another
  - All data points are from the same underlying distribution

- Most parametric tests for numeric data assume the data is normally-distributed
  - Or, sample size is large enough to "waive" this assumption, thanks Central Limit Theorem!!

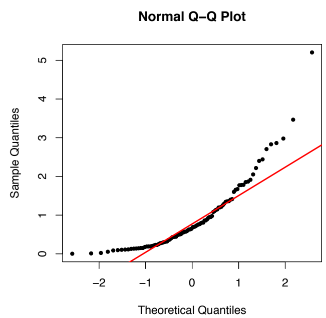
Assessing normality with a Q-Q plot

Quantile-**Q**uantile plots graphically show if two datasets come from the same distribution

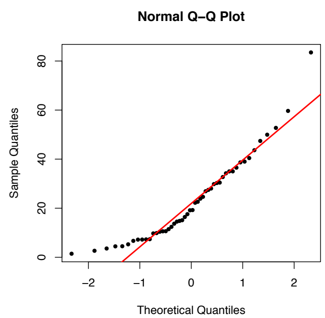
- If the points follow the "expectation" line, datasets are similarly distributed



Ideal scenario. Data is normal

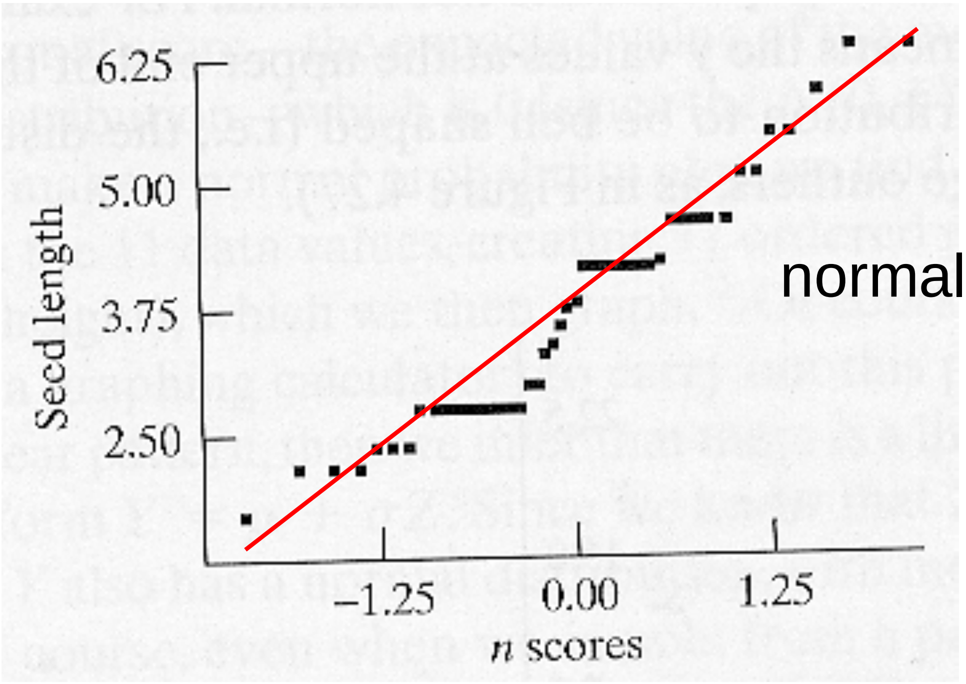


Data is not normal



Close enough, let's say normal

Granular data is also normal!

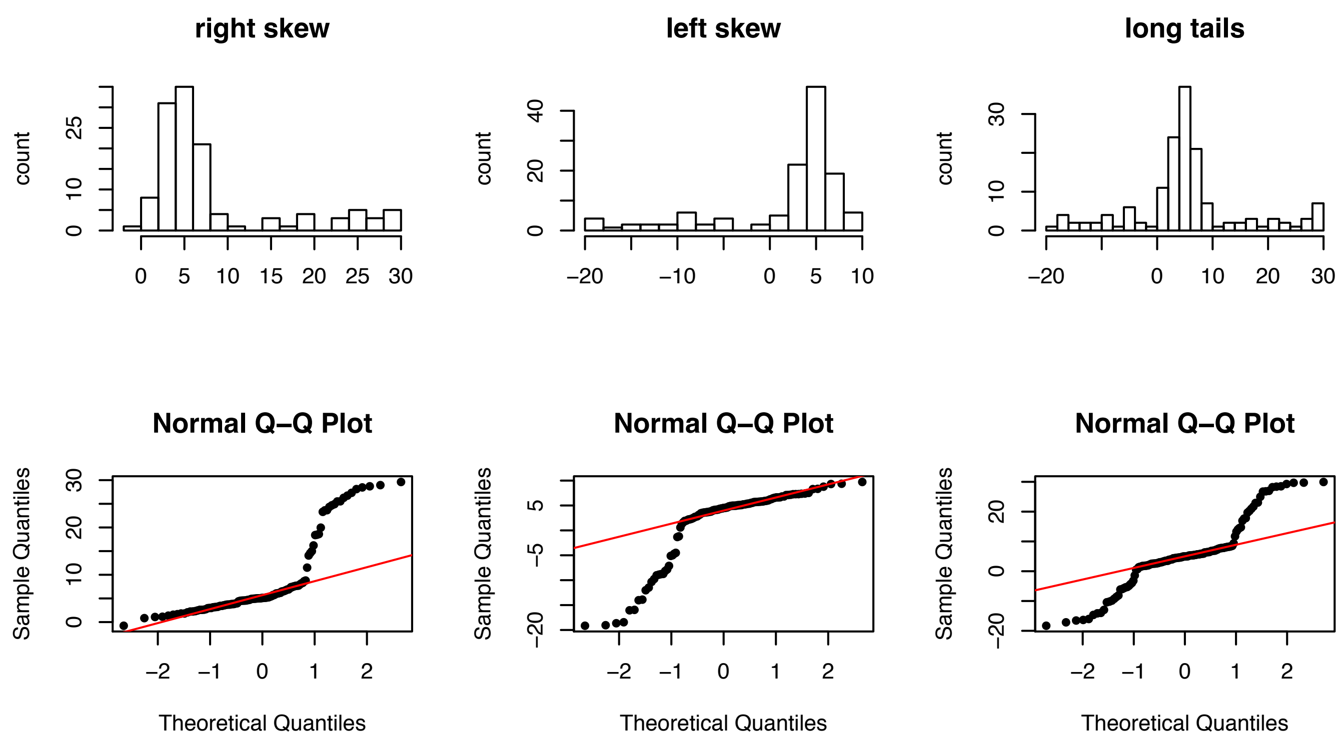


Troubleshooting: Failure to meet normal assumption

- 1. If sample size is large enough (>~30), Central Limit Theorem kicks in and assumptions are effectively met

- 2. If sample is size is small (<~30) we can either:
  - **Transform** the data to be normal
  - Use a **nonparametric test**

Non-normal data



Data transformations

- Log of data: x → log(x)
- Square root of data: x → sqrt(x)
- Inverse of data: x→ 1/x

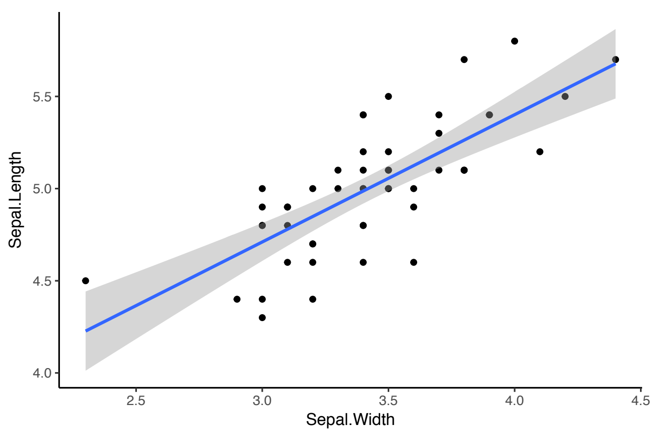
Data transforms: Caution

- Your test will now run on **transformed data**
  - Assume log transform performed and result has effect size 1.5
  - **Actual effect size is exp(1.5) = 4.48**

- Be careful of 0's in data
  - 1/0 and log(0) are undefined
  - Hack: Replace all 0's with tiny number like 1e-8

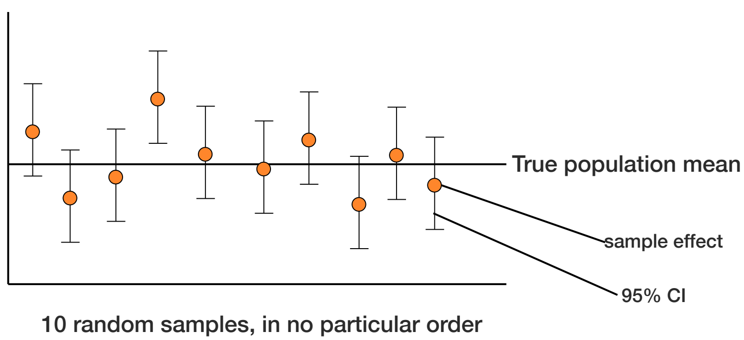
We’re doing linear models (regression and beyond!)

- These tests use **the *t* distribution** to assess significance



Confidence intervals

- Range of values surrounding the sample estimate that is likely to contain the population parameter
- If I took 100 random samples, 95 of their 95% CIs would contain the true parameter
  - If I took 100 random samples, 50 of their 50% CIs would contain the true parameter… etc.



9/10 (~95%) of these random samples has a 95% confidence interval that overlaps the true mean

(blank page for Q/A?)

(blank page for Q/A?)