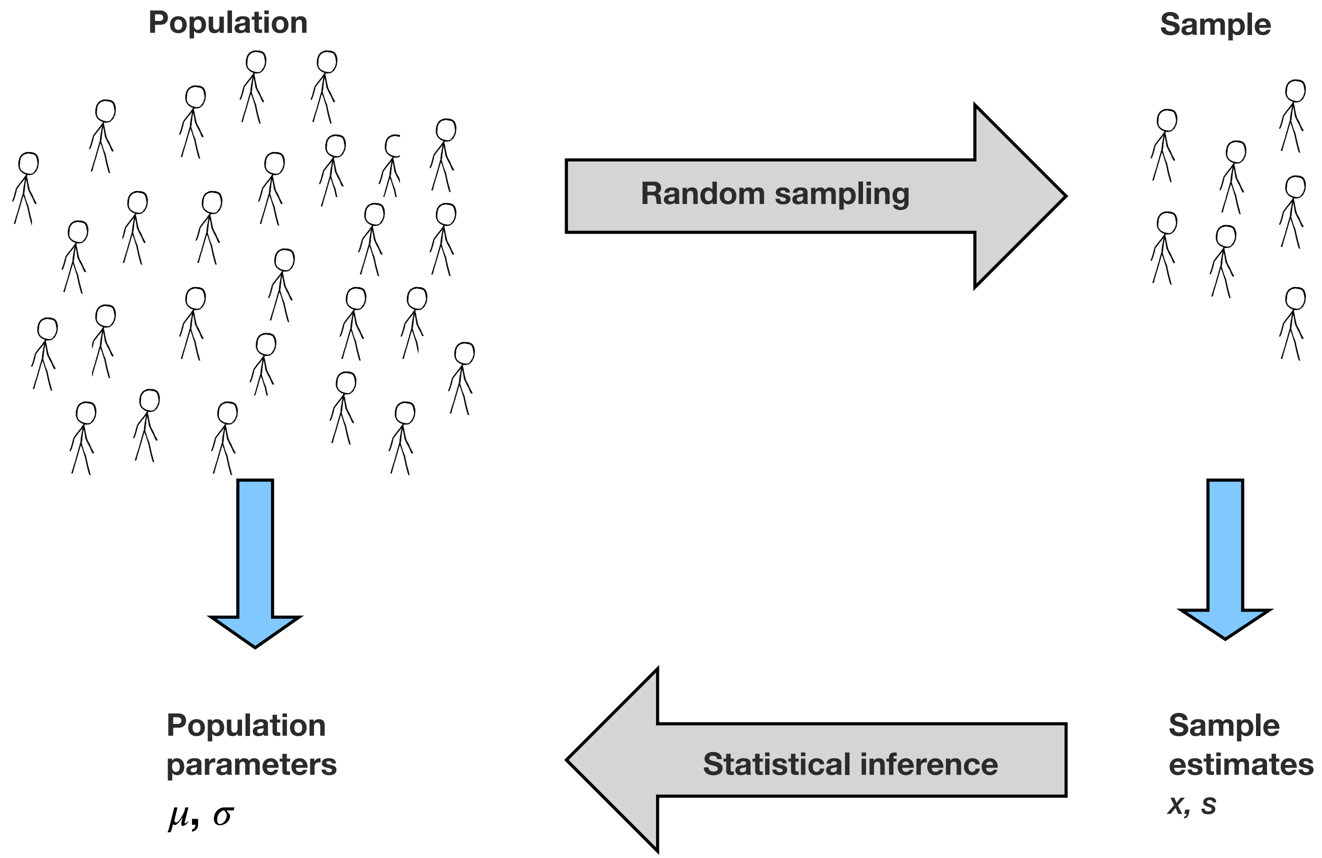
Hypothesis Testing I: Introduction and Comparing means

BIO5312 FALL2017

STEPHANIE J. SPIELMAN, PHD

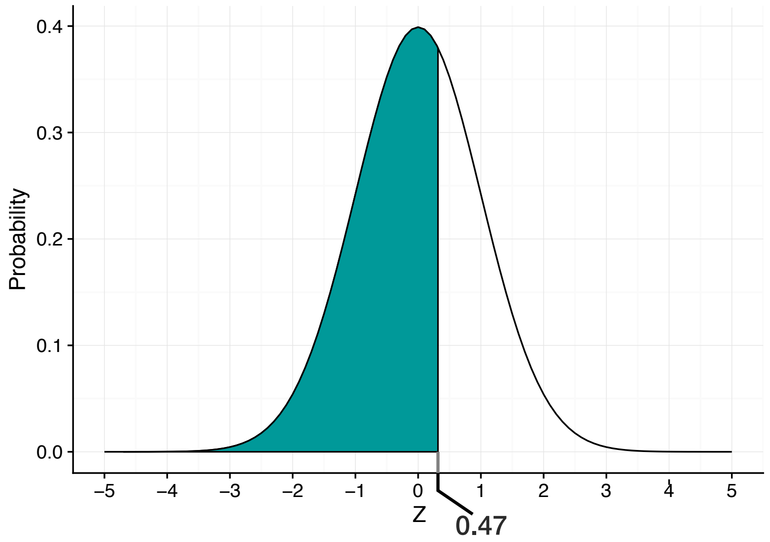
Briefly, estimation



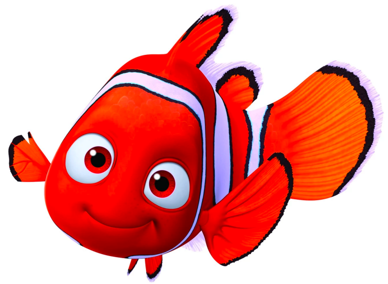
Recall from last week

For $X \sim N(0,1)$, what is the probability P(X ≤ 0.47)?

```
# CDF: P(X <= 0.47)
> pnorm(0.47)
   [1]  0.6808225
```



Computing probabilities for a population



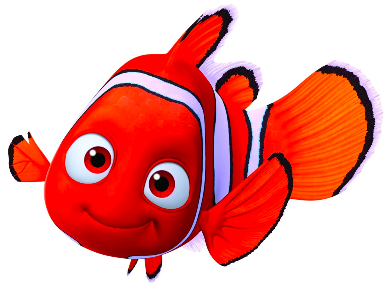
Clownfish lengths are normally distributed as N(11, 1.9).

What is the probability that a clownfish is less than 13 cm?

$$Z = \frac{x-\mu}{\sigma} = \frac{13-11}{\sqrt{1.19}} = 1.83$$

```
> pnorm(1.83)
  [1] 0.966375
```

Computing probabilities for a sample



Clownfish lengths are normally distributed as N(11, 1.9).

What is the probability that a sample of 100 clownfish has a mean less than 11.5?

$SE = \frac{\sigma}{\sqrt{n}} = \frac{11}{\sqrt{100}} = 1.1$ → The standard deviation of the sampling distribution

$$Z = \frac{x-\mu}{SE} = \frac{11.5-11}{1.1} = 0.454$$

```
> pnorm(0.454)
[1] 0.6750856
```

A return to sampling distributions

The **sampling distribution** is the distribution of all possible values an estimate for a parameter can take

- Sampling distribution of the means = all the means one could measure across samples

DEMO: Part 3 on Homework

Hypothesis testing

Compare data (random sample) to the expectation of a specific null hypothesis

Hypothesis testing uses probability to answer whether an observed effect occurred by chance

Example scenario to use hypothesis testing

The polio vaccine was first tested in 1954.

~400,000 students were divided into two random groups:

- Half received the vaccine, half received placebo
- In vaccine group, 0.016% developed polio.
- In placebo group, 0.057% developed polio.

We can use hypothesis testing to ask if the vaccine likely worked, or whether random chance likely caused results.

Hypothesis testing has null and alternative hypotheses

The **null hypothesis** H_0 makes a claim about the underlying population parameter

- "Nothing interesting is going on"
- Specific tests have specific null hypotheses

The **alternative hypothesis** H_A is what we would like to "know if it's true"

- "Something we care about is going on"

Parametric vs. nonparametric hypothesis tests

Parametric tests assume that the data follow a particular known distribution

- Distributions such as normal, binomial, chi-squared, etc.

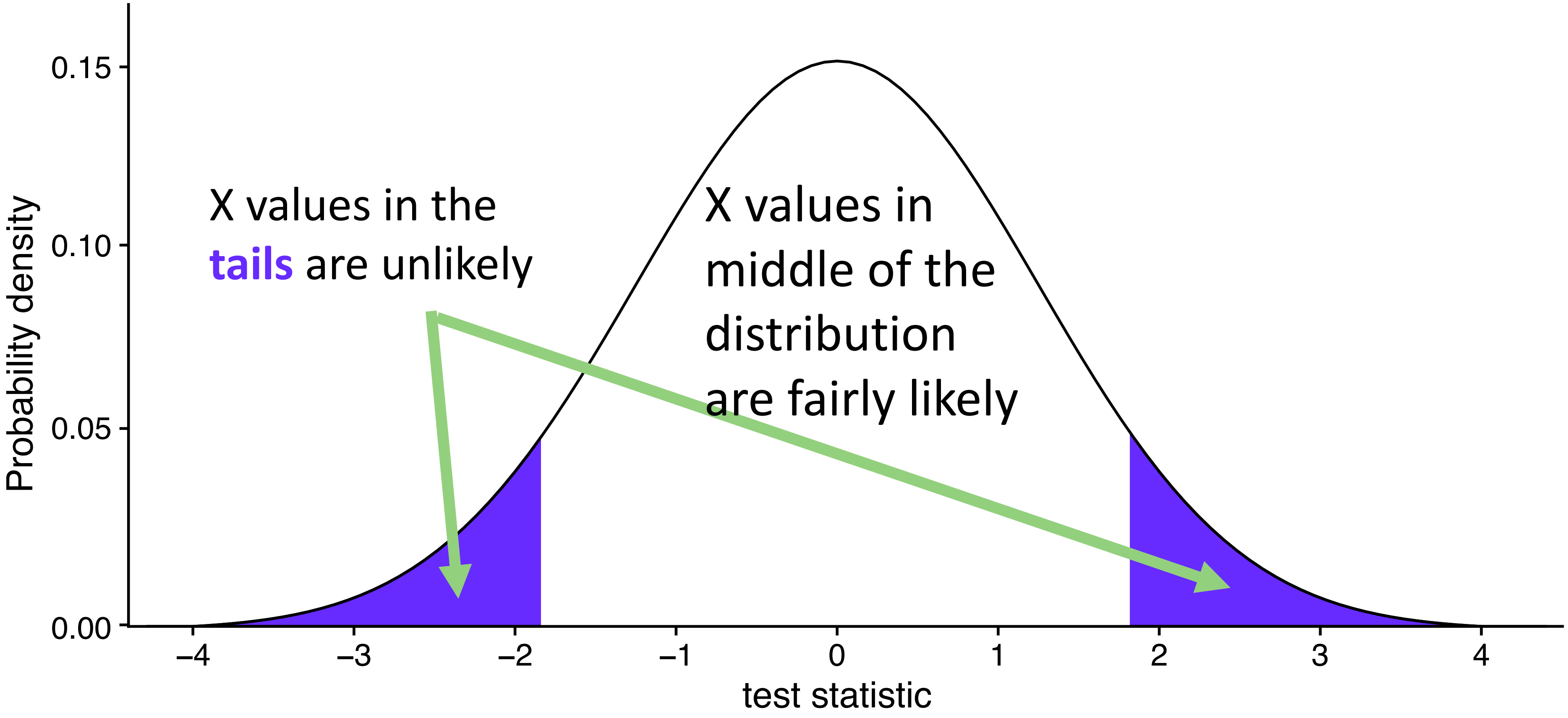
Nonparametric tests make no assumption about the data and are "distribution-free"

The null distribution represents H_0

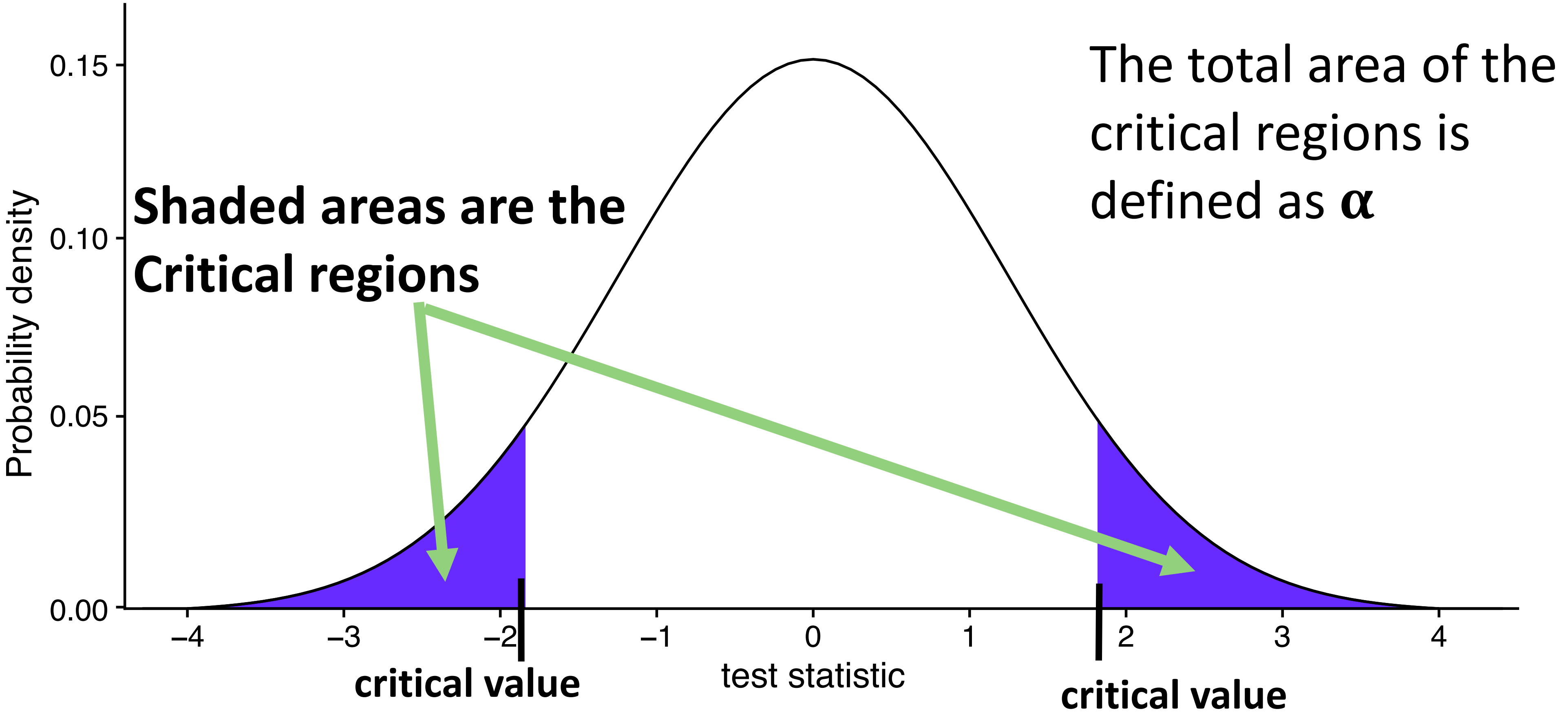
The *null distribution* is the sampling distribution of outcomes for a test statistic assuming the null is true

Hypothesis tests ask:
To what extent are my data expected under the null hypothesis?

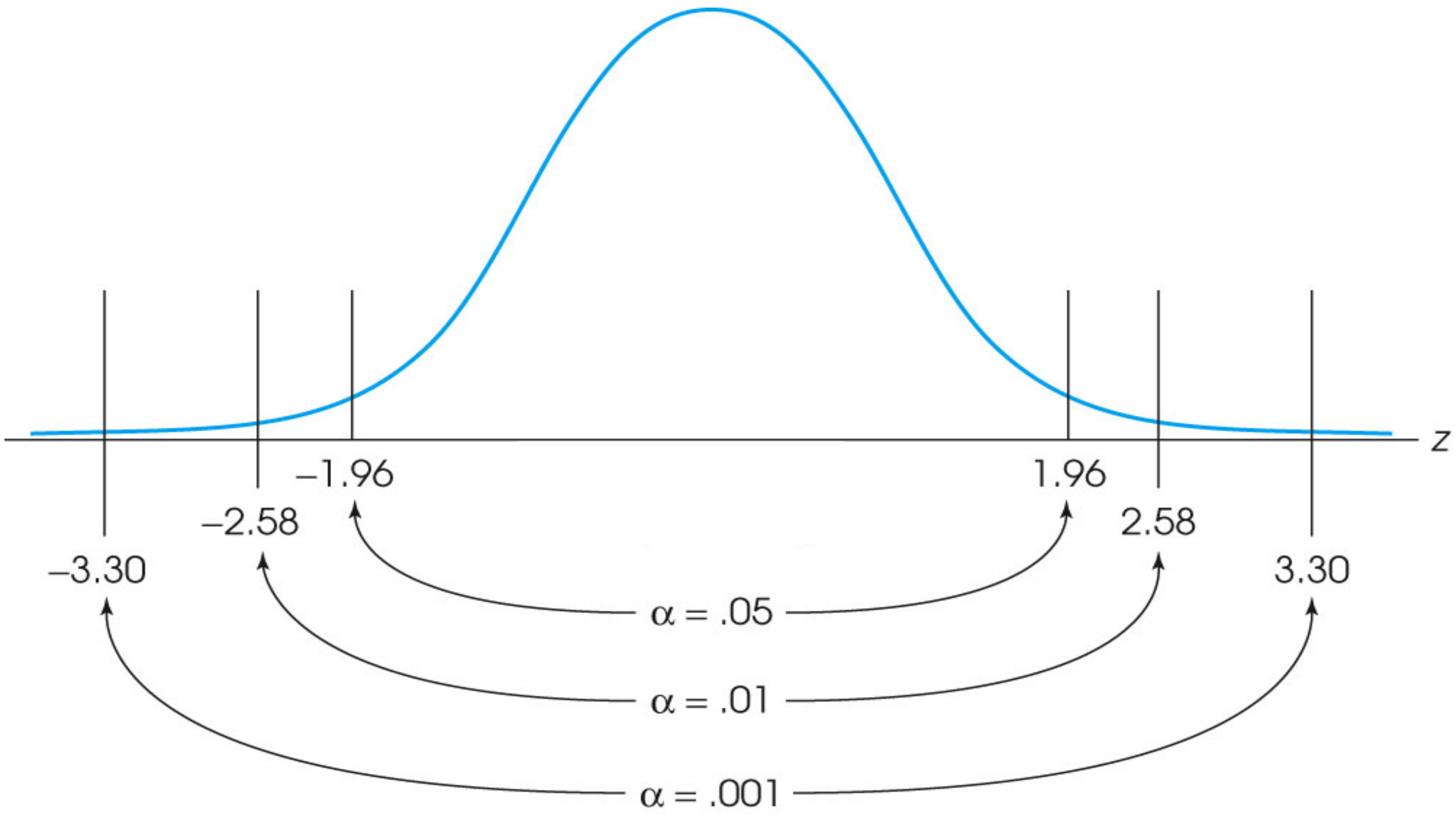
Sampling distribution of the null hypothesis



Sampling distribution of the null hypothesis

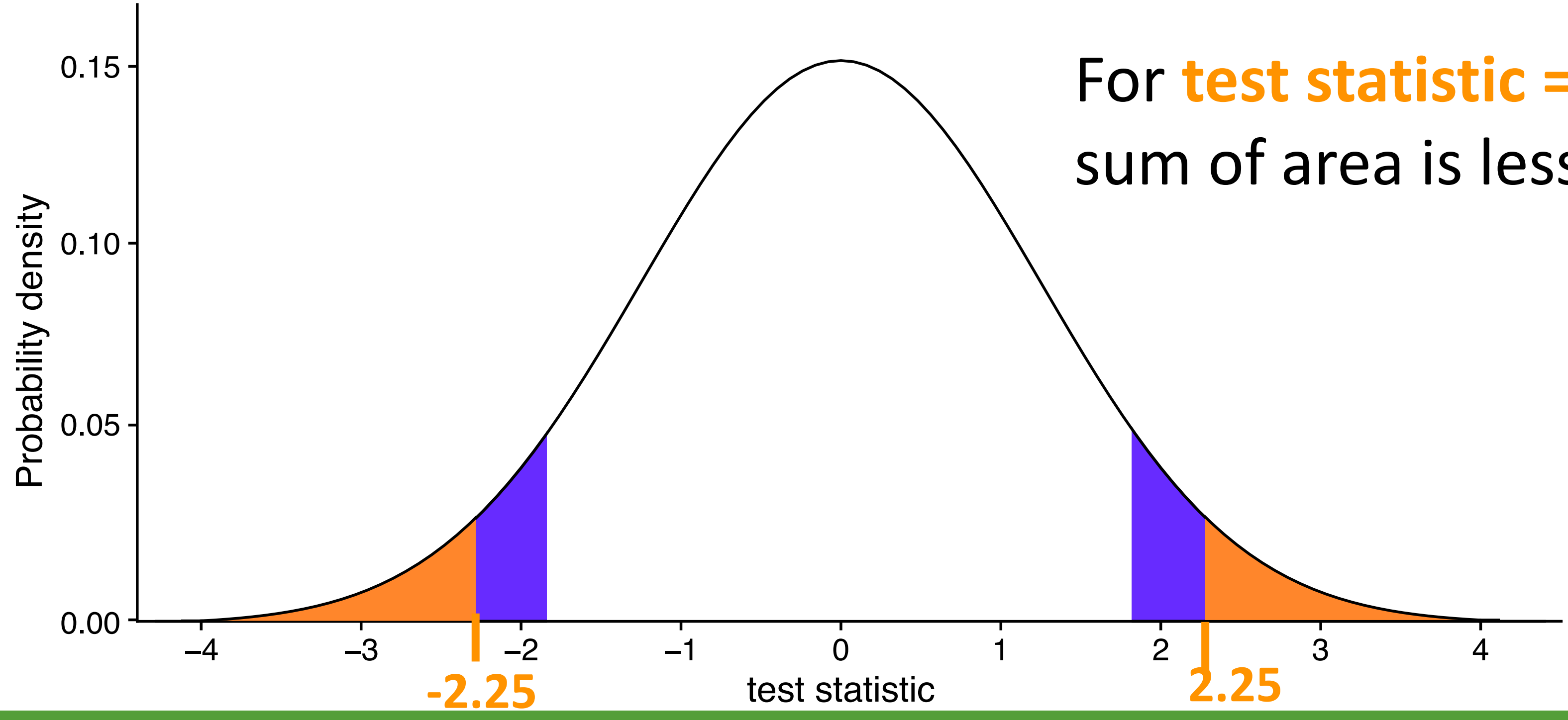


Critical values where N(0,1) is the null



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



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

NOTE: It will not always be symmetric

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

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 α=0.05

Truth about population (generally unknown)

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

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. **FN (type II)**

A person with HIV receives a positive test result for HIV. **No error**

A person using illegal performing enhancing drugs passes a test clearing them of drug use. **FN (type II)**

A study found a significant relationship between neck strain and jogging, when reality there is no relationship. **FP (type I)**

An healthy individual gets a positive cancer biopsy result. **FP (type I)**

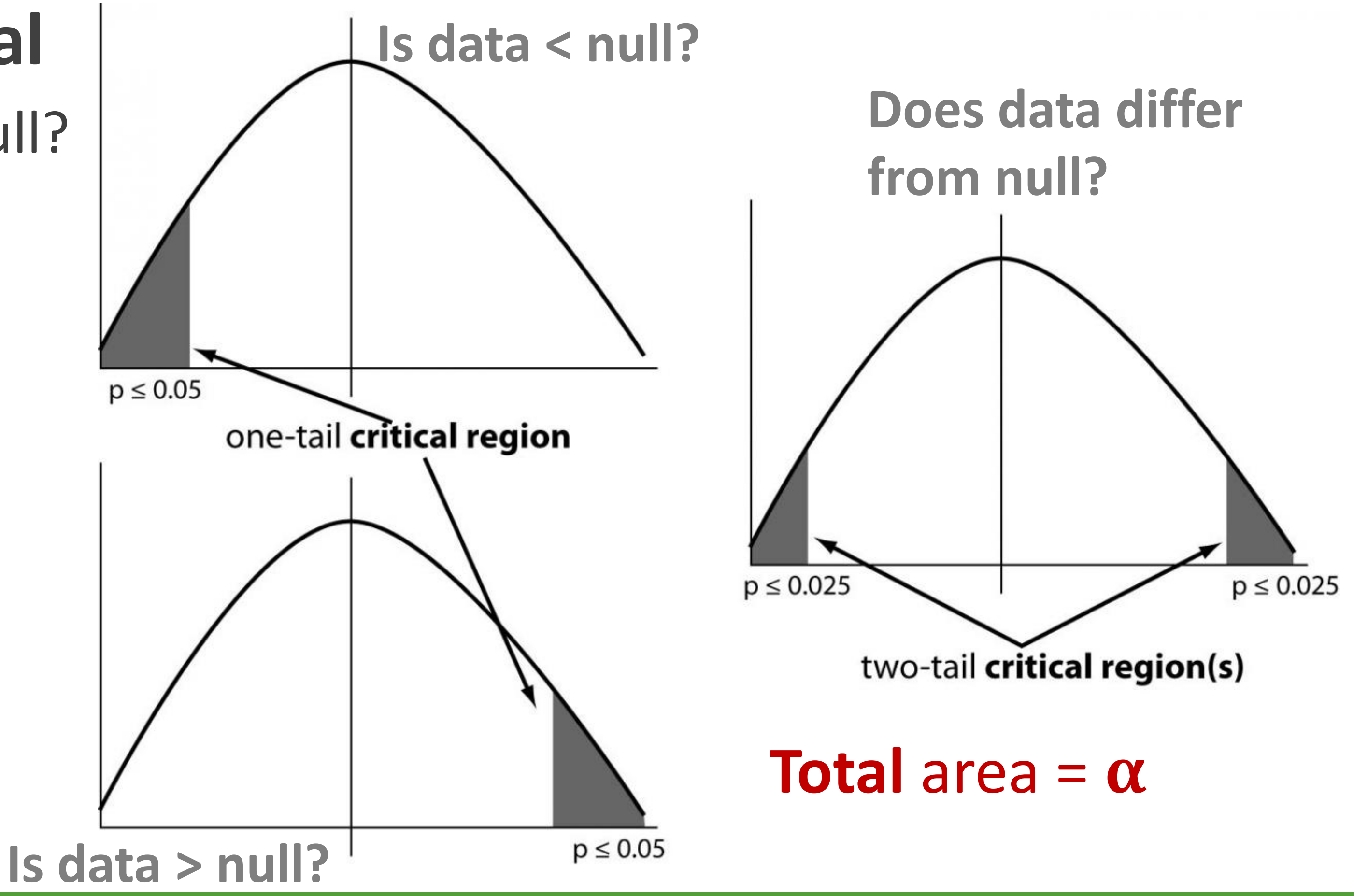
One-sided vs. Two-sided (or –tailed)

One-sided tests are **directional**

- Are my data larger/smaller than null?

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

- Do my data differ from null?



Total area = α

One-sided vs. Two-sided tests

One-sided tests have **more power** than two-sided tests

- Power = the ability to detect a true effect
- Also known as **true positive rate**

One-sided tests are more limited in scope and can get you in trouble if you choose the wrong direction

You must choose only one test before you look at your data

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 $\boldsymbol{\alpha}$ (P-value threshold)

*Parametric only

Hypothesis tests to compare means

Test the null hypothesis:

- **One sample test:** The mean of my sample equals a null value
- **Two sample test:** The mean of my two samples are equal (difference in means is 0)

Two options:

- **Z-test**, where the null distribution is the standard normal N(0,1)
  - Test statistic is **Z**
- ***t*-test**, where the null distribution is the Student's *t* distribution
  - Test statistic is ***t***

Z-test vs *t*-test

$Z = \frac{x-\mu}{\sigma}$ for populations

$Z = \frac{x-\mu}{SE}$ where SE = $\sigma / \sqrt{n}$ for samples

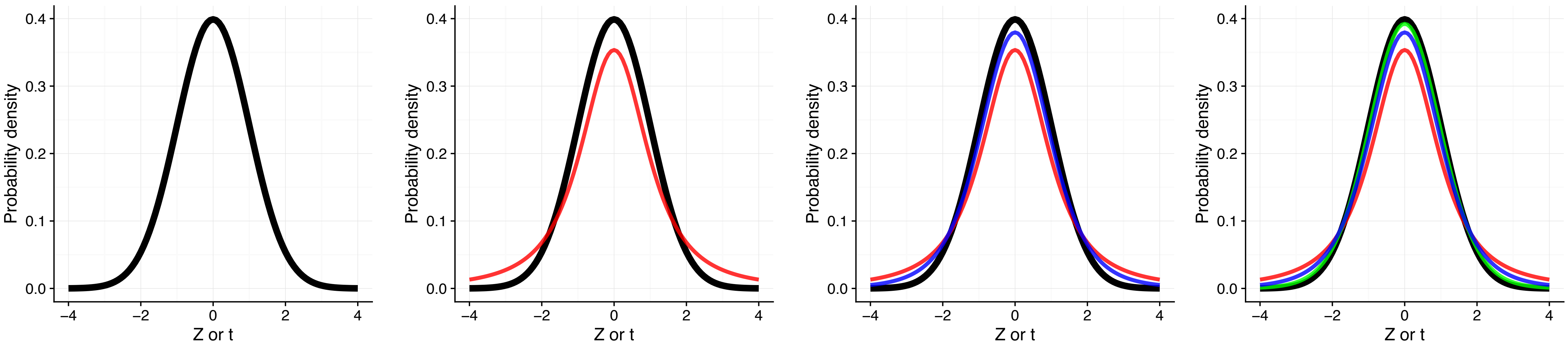
Because SE is rarely known, we can approximate with $\frac{s}{\sqrt{n}}$, leading to *t*-tests

$$t = \frac{\bar{x}-\mu}{SE_{\overline{x}}}, \quad \text{where } SE_{\bar{x}} = \frac{s}{\sqrt{n}}$$

Standard normal and Student's *t*

Adding *t* distributions with increasing *degrees of freedom*

Standard normal N(0,1)

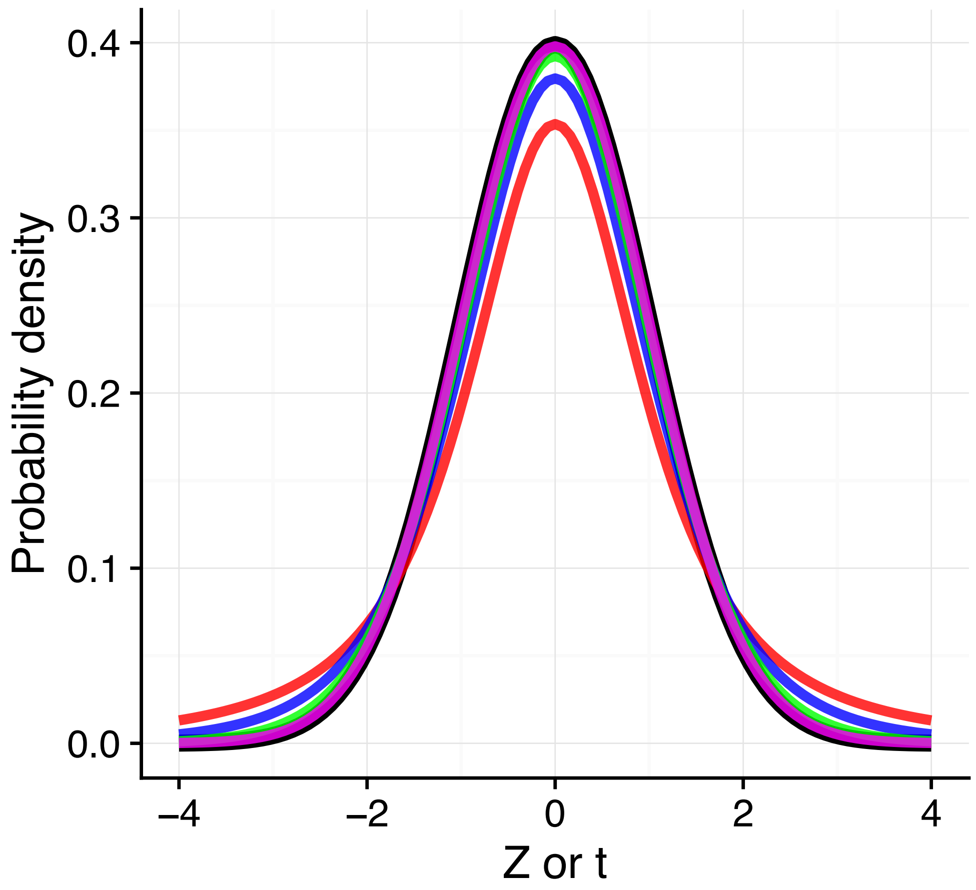


df = 2

df = 5

df = 15

Student's *t* where df = 100



Properties of the *t* distribution

As df approaches ∞, the *t* distribution approaches N(0,1)

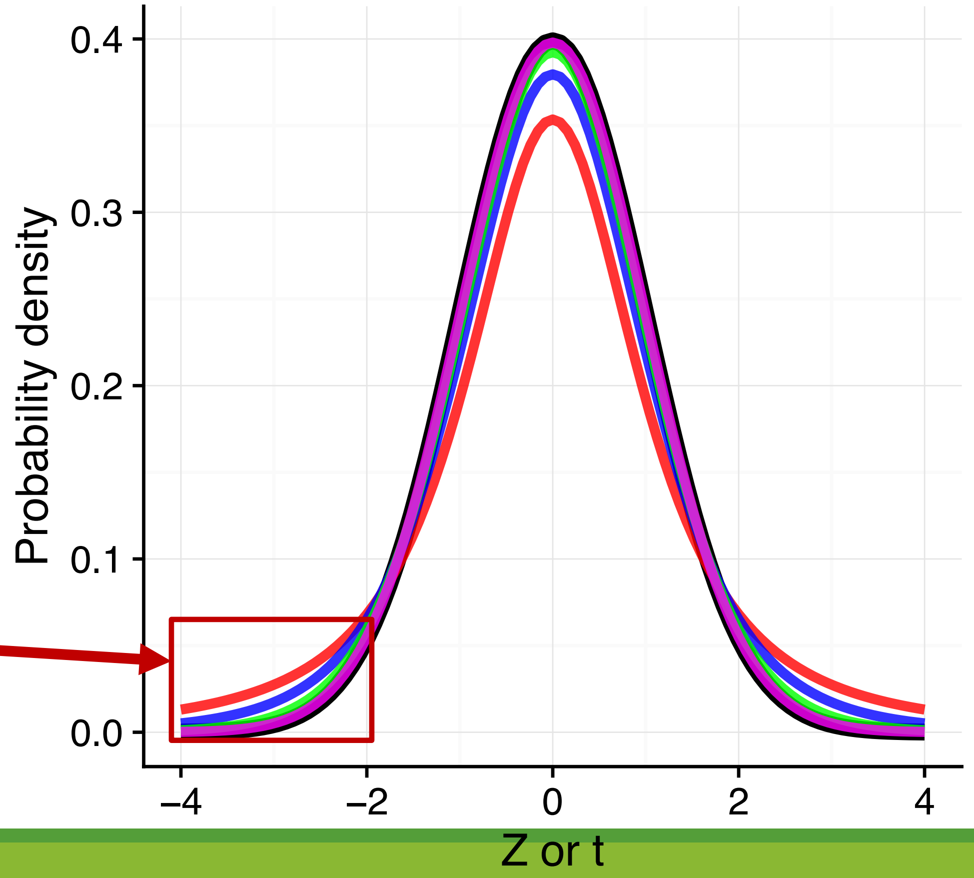
- Usually, df = n – 1, where n is sample size

Like the normal distribution…

- *t* distribution is symmetric
- Mean = median = mode

Unlike the normal distribution…

- *t* has *much fatter tails*



Test Assumptions

t-tests assume data is normally distributed

- This applies to **both samples** for a two-sample *t*-test

Dive right in: One sample *t*-test

Null hypothesis

$\mathbf{H_0 : \bar{x} = \mu}$

One-sided test

$\mathbf{H_A : \bar{x} > \mu}$

$\mathbf{H_A : \bar{x} < \mu}$

Directional

Two-sided test

$\mathbf{H_A : \bar{x} \neq \mu}$

Non-directional

Performing a one-sample *t*-test

I want to know if the disease, Bad Disease, influences human body temperature. We know that standard human body temperature is **98.6** degrees F. I measured the temperatures for a **random sample of 15 individuals** with Bad Disease. On average, they have a temp of **99.59** degrees F.

Does Bad Disease raise body temperature?

H_0 : Bad Disease does not raise body temperature. $\bar{x} = 98.6$

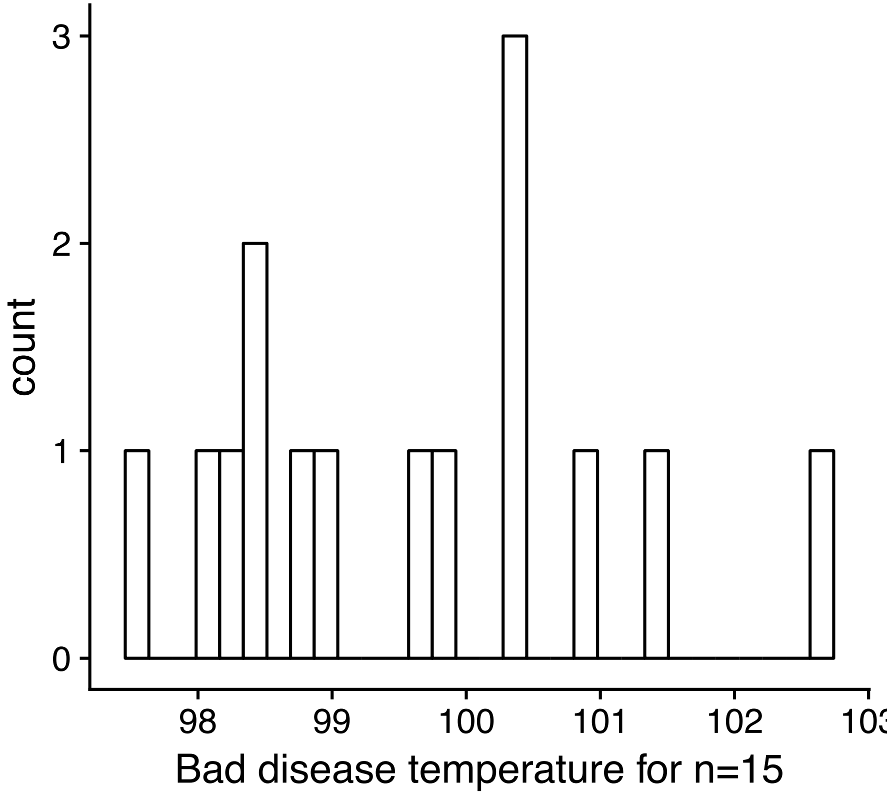
H_A : Bad Disease raises body temperature. $\bar{x} > 98.6$

```
> head(bad.disease.temp)
     temp
1  98.17420
2  97.62137
3  99.60920
4 100.44158
5  99.75483
6 100.28846


> mean(bad.disease.temp$temp)
  [1] 99.594
```

Checking assumptions of the test

The *t*-test assumes that our sample data is normally distributed

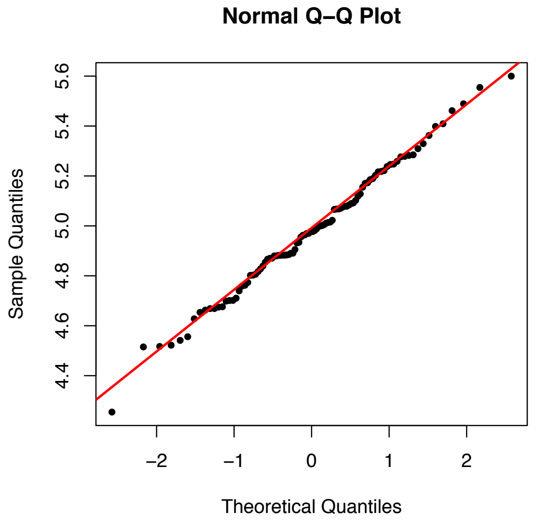


Hard to tell if normal from histogram!

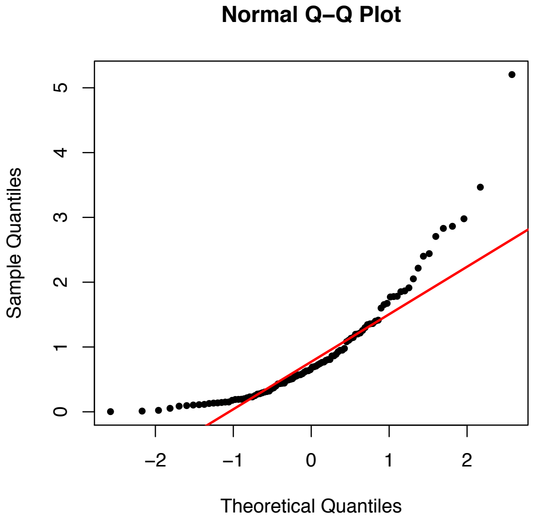
Assessing normality with a Q-Q plot

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

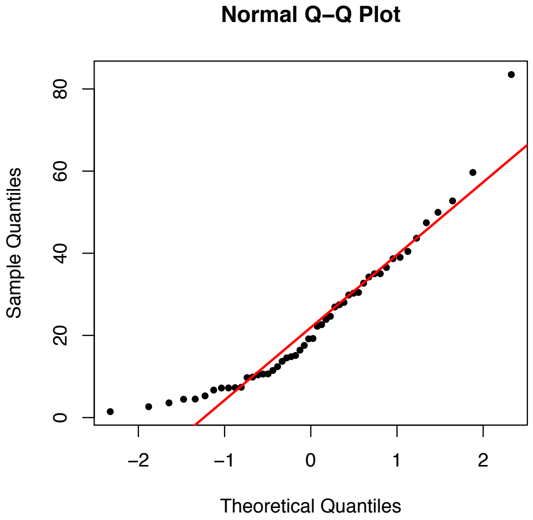
- If the points follow the "expectation" line, datasets are similarly distributed



Ideal scenario. Data is normal

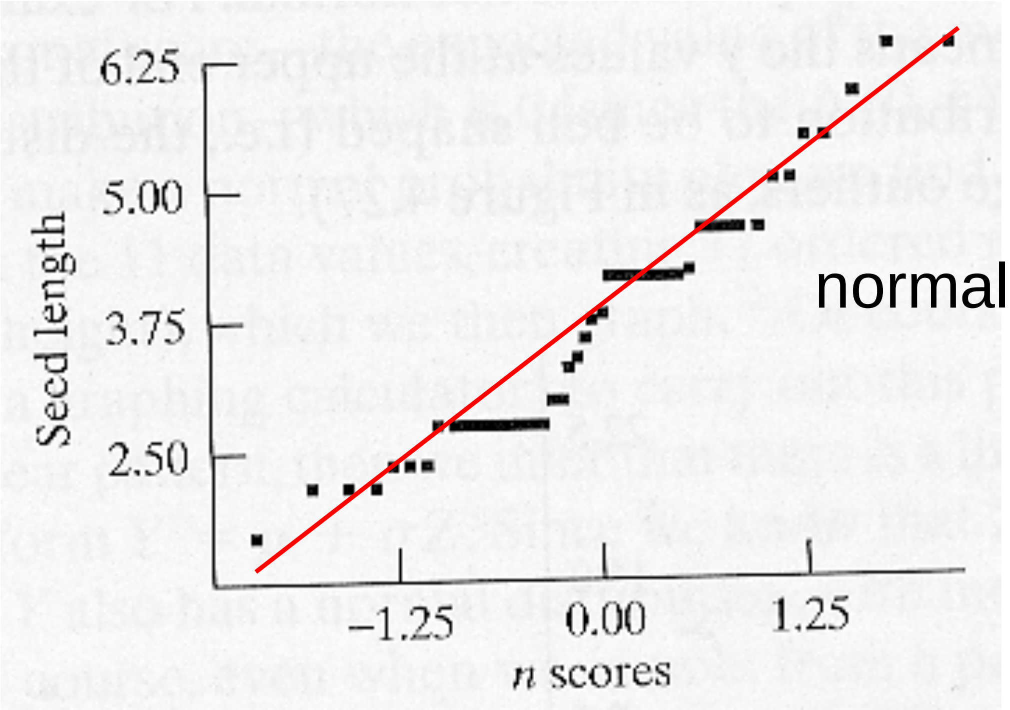


Data is not normal



Close enough, let's say normal

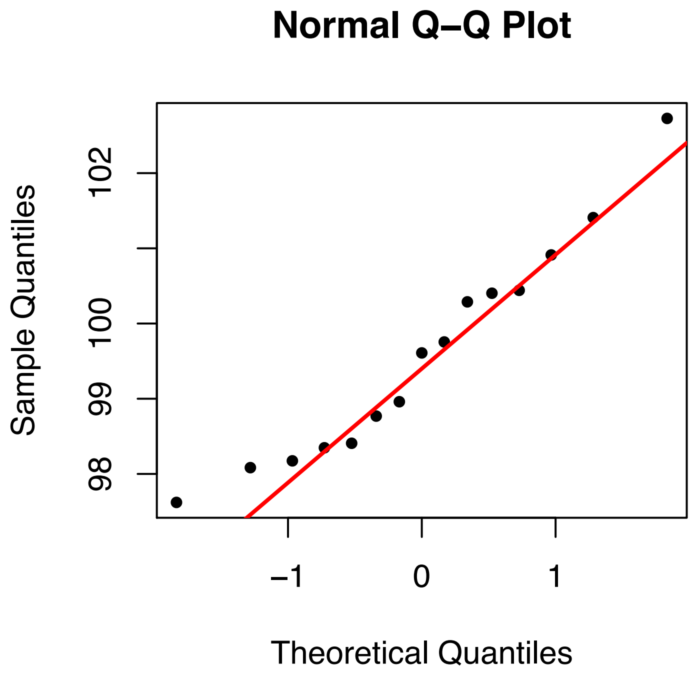
Granular data is also normal!



Making a Normal Q-Q plot

```
> head(bad.disease.temp)
     temp
1  98.17420
2  97.62137
3  99.60920
4 100.44158
5  99.75483
6 100.28846

## Uses base R, not ggplot ##
> qqnorm(bad.disease.temp$temp, pch=20)
> qqline(bad.disease.temp$temp, col="red")
```



Data approximately follow the QQ line. Therefore, assumptions have been met and we can run the test.

Performing the *t*-test

1. Calculate the *test statistic*
2. See where test statistic falls on its distribution
3. Compute P-value as **area under the curve** past this statistic
    - The P-value is the probability of obtaining a test statistic as large or larger than that recovered

Compute the test statistic, *t*

H_0 : Bad Disease does not raise body temperature. $\bar{x} = 98.6$
H_A : Bad Disease raises body temperature. $\bar{x} > 98.6$

```
> mean(bad.disease.temp$temp)
  [1] 99.594
> sd(bad.disease.temp$temp)
  [1] 1.438273
> nrow(bad.disease.temp)
  [1] 15
```

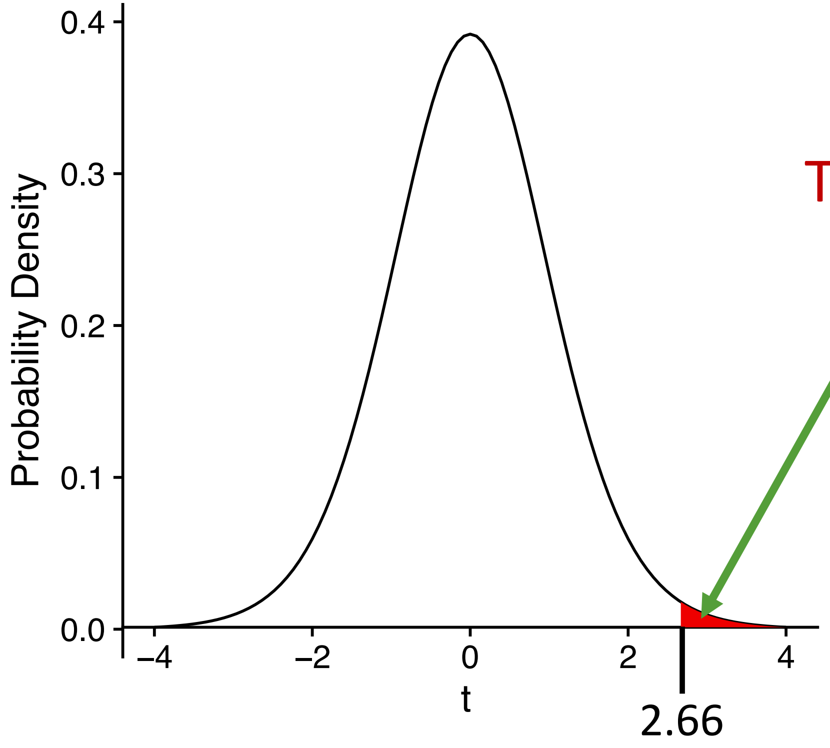
$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} = \frac{99.59 - 98.6}{\frac{1.44}{\sqrt{15}}} = 2.66$$

this is our test statistic, $t_{2.66}$

More precisely, $t_{2.66,\ df=14}$

Find where the statistic falls in distribution

Our null distribution is a *t* distribution with df = 14 (15-1)



This area is our **P-value** because we test if $\bar{x} \boxed{>} \mathbf{98.6}$

```
## P(X >= 2.66) ##
> 1 – pt(2.66, 14)
   0.009
```

Forming conclusions

Our P=0.009, which is less than α=0.05. Therefore we reject the null hypothesis and we have evidence that Bad Disease raises body temperature.

H_0 : Bad Disease does not raise body temperature. $\boldsymbol{\bar{x} = 98.6}$
H_A : Bad Disease raises body temperature. $\boldsymbol{\bar{x} > 98.6}$

Approach to hypothesis testing

1. Decide what question you are interested in answering

   Is the mean of my data equal to 98.6?

2. Determine the appropriate hypothesis test to use

   Use a one-sample *t*-test

3. Check that your data meet the assumptions of the test

   We confirmed the data is normally distributed

4. Compute the *test statistic* for your hypothesis test and the corresponding P-value

   We found t = 2.66 and P=0.009

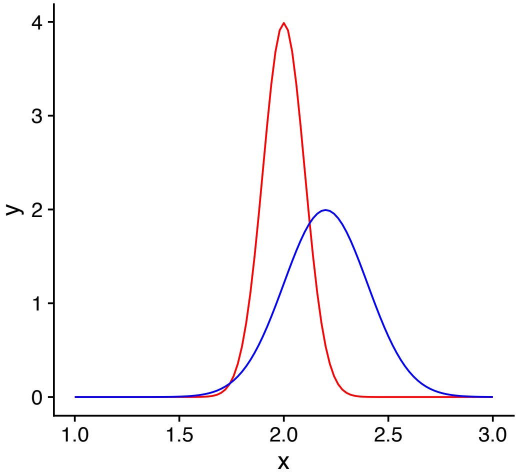
5. Draw conclusions using a specified P-value threshold

   At α = 0.05, we reject the null hypothesis and find evidence that Bad Disease raises body temp.

Effect size

The effect we observed here is **99.59 – 98.6 = 0.99**

Statistical significance is **not the same** as biological significance



Comparison between red and blue samples will show a significant difference between means.
But does it matter?

Confidence intervals

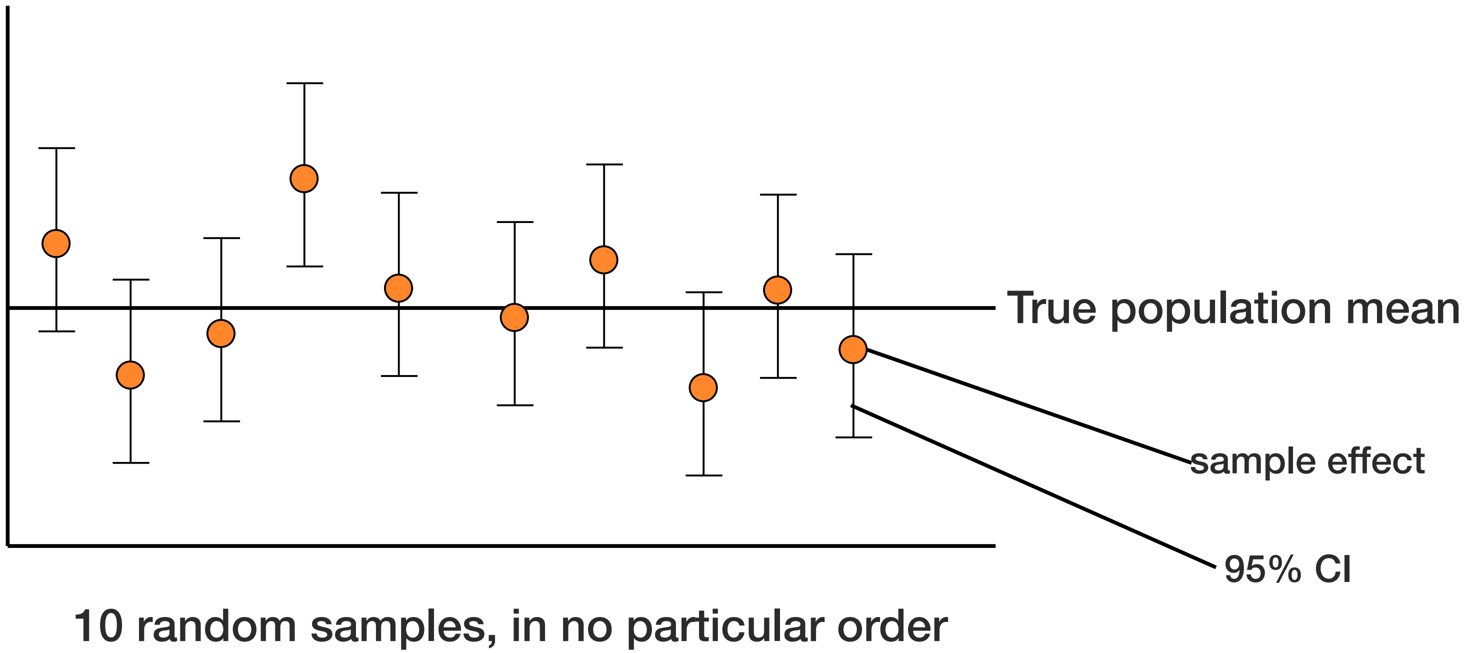
Range of values surrounding the sample estimate that is likely to contain the population parameter

Generally we calculate the **95% confidence interval** (goes with $\boldsymbol{\alpha}$ = 0.05)

In 95% of random samples, this will be true:

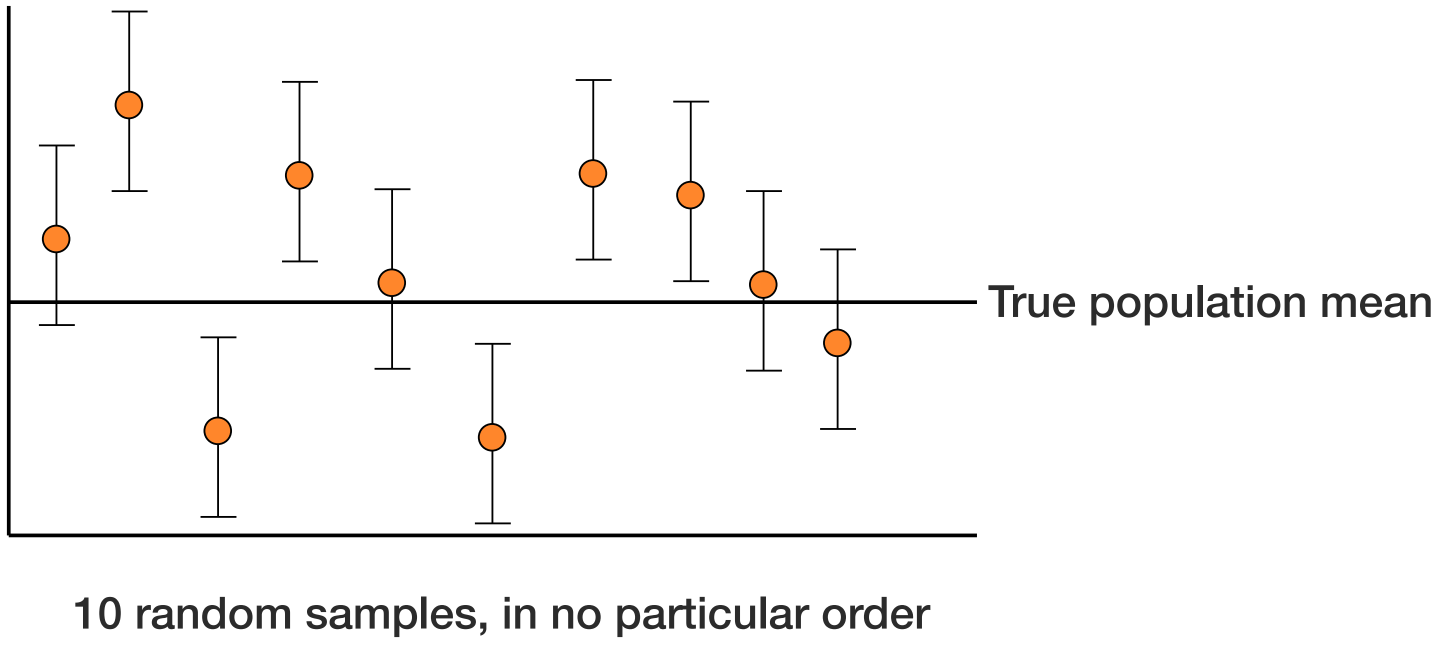
$$\boldsymbol{\bar{x} - (t_{0.025} * SE_{\bar{x}}) < \mu < \bar{x} + (t_{0.025} * SE_{\bar{x}})}$$

Conceptualizing the CI



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

Does this figure represent 95% CIs?



NO. If anything, these are 40% CIs

Confidence intervals

Construct upper and lower limits of 95% CI

- Lower: $\bar{x} - (t_{0.025} * SE_{\bar{x}})$
- Upper: $\bar{x} + (t_{0.025} * SE_{\bar{x}}\ )$

95% CI = $\boldsymbol{\bar{x} \pm (t_{0.025} * SE_{\bar{x}}\ )}$

$$\boldsymbol{= 99.59 \pm 0.189}$$

The true population mean is 95% likely to be in the range 99.4 – 99.78

```
### Calculate t_0.025
> pt(0.025, 14)
[1] 0.5097961

### Calculate t_0.025*SE
> t <- pt(0.025, 14)
> se <- sd(bad.disease.temp$temp)/sqrt(15)
> t * se
   [1] 0.1893181
```

Bring it all together

We have a sample mean of 99.59 with a standard error of 0.19.

Our test statistic $t_{df=14}$ = 2.66, giving a P-value = 0.009. We reject the null hypothesis at α = 0.05 and have evidence that Bad Disease raises temperature.

Our effect size is 0.99.

We found a 95% CI of 99.59 $\pm$ 0.189, giving the likely range for the true population parameter. Note that the null of 0 is not in the CI.

Reporting non-significant results

Let's say we found $t_{df=14} = 1.05 \rightarrow P = 0.15$

```
### Calculate P-value
> 1 - pt(1.05, 14)
[1] 0.1557531
```

At **α** = 0.05, we **fail to reject** the null hypothesis. We have no evidence that Bad Disease raises body temperature.

- Does **not** mean that Bad Disease doesn't raise the temperature – our sample just had no evidence for this effect.

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 | H_0 is true)**
- **P-value = 0.009:** If H_0 is true, I would obtain this effect or larger ($t>=2.66$) in 0.9% of such studies due to random sampling error

What **is** a P-value?

A low P-value means that the data are **unlikely under the null**

- We therefore make an educated guess that there is probably something else going on, such as the alternative hypothesis
- We **can never** rule out the possibility that results were fully consistent with null, just unlikely

P-values **are not magic**

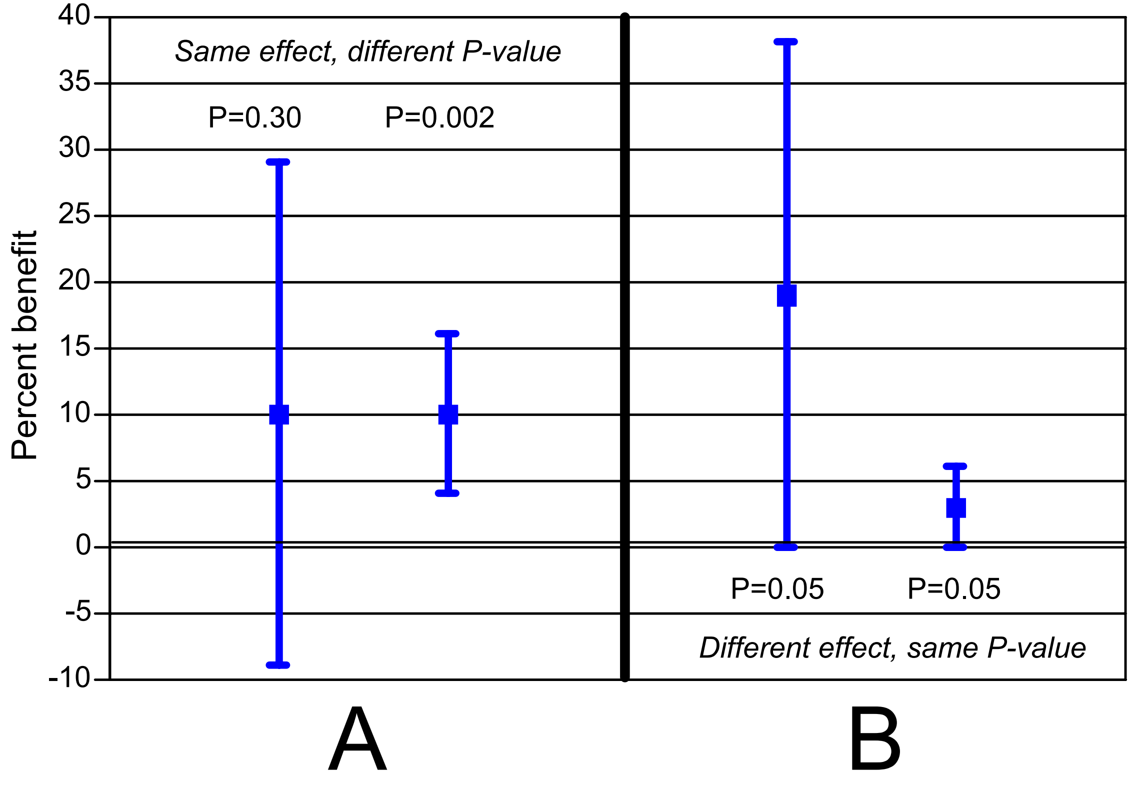
P-values **cannot** evaluate whether H_0 or H_A is true

- 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 likely isn't.

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 and effect size



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 sampling 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

To recap

We have performed a **one-sided *t*-test** to test the alternative hypothesis that $\bar{\boldsymbol{x}} > \mathbf{98.6}$

We can also perform a **two-sided *t*-test** to test the *non-directional* alternative hypothesis that $\bar{\boldsymbol{x}} \neq \mathbf{98.6}$

Compute the test statistic, *t*

H_0 : Bad Disease does not affect body temperature. $\bar{x} = 98.6$
H_A : Bad Disease affects body temperature. $\bar{x} \neq 98.6$

$\bar{x}$ = 99.59
μ = 98.6
n = 15

```
> head(bad.disease.temp)
     temp
1  98.17420
2  97.62137
3  99.60920
4 100.44158
5  99.75483
6 100.28846

> mean(bad.disease.temp$temp)
   [1] 99.594
> sd(bad.disease.temp$temp)
   [1] 1.438273
```

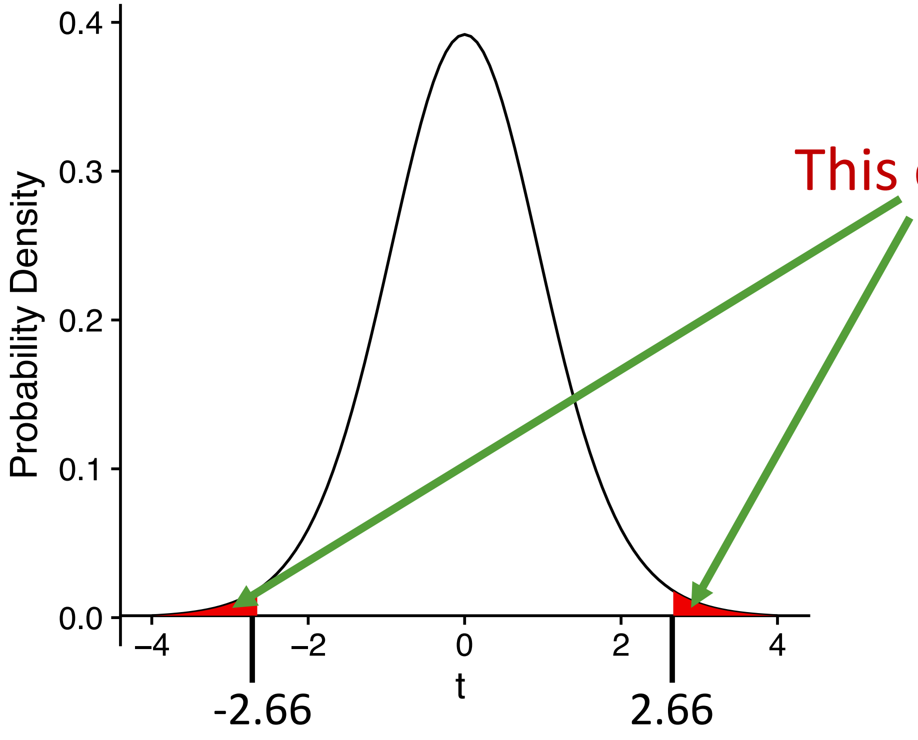
$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} = \frac{99.59 - 98.6}{\frac{1.44}{\sqrt{15}}} = 2.66$$

this is our test statistic, $t_{2.66}$

More precisely, $t_{2.66,df=14}$

Computing the P-value

For a two-sided test, we consider **both extremes**



This combined area is our **P-value = 0.018**

```
## P(X >= 2.66) or P(X<=-2.66)
> 2 * (1 - pt(2.66, 14))
   0.01806
```

When running a one-sided *t*-test goes wrong

For sample of n=20, I want to test $\bar{x} < \mu$ where μ = 5.8

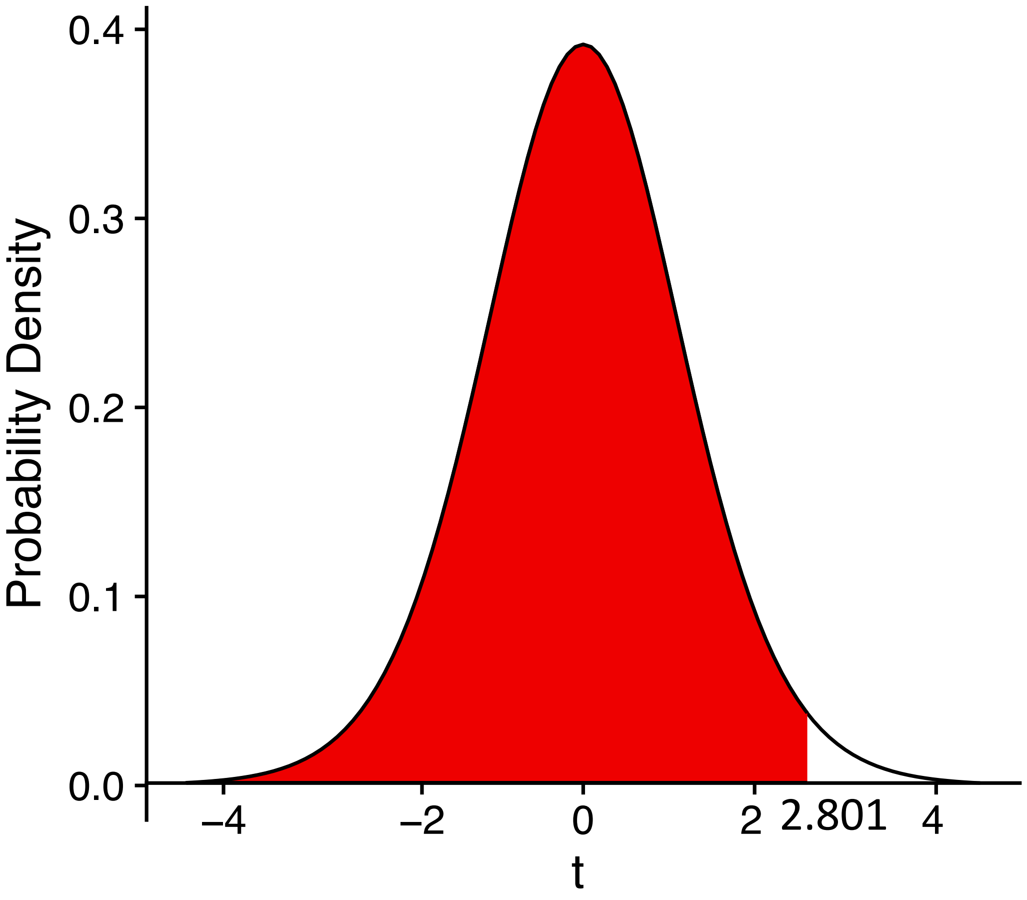
```
> head(example.sample)
     values
 1 5.511307
 2 5.012612
 3 6.178421
 4 7.587568
 5 6.892165
 6 5.197049

> mean(example.sample$values)
  [1] 6.325366
> sd(example.sample$values)
  [1] 0.8295474
```

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} = \frac{6.32 - 5.8}{\frac{0.83}{\sqrt{20}}} = 2.801$$

The area **below** the statistic is our P-value for $\bar{x} < \mu$

```
> pt(2.801, 19)
[1] 0.9943006
```



Two-sample *t*-tests

Compare means of **two samples** ($\bar{x}_1$ and $\bar{x}_2$) to each other.

Are the underlying population means μ_1 and μ_2 the same?

Null hypothesis

$H_0 : \mu_1 = \mu_2$

$H_0 : \mu_1 - \mu_2 = 0$

One-sided test

$H_A : \mu_1 > \mu_2$

$H_A : \mu_1 < \mu_2$

$H_A : \mu_1 - \mu_2 > 0$

$H_A : \mu_1 - \mu_2 < 0$

Two-sided test

$H_A : \mu_1 \neq \mu_2$

$H_A : \mu_1 - \mu_2 \neq 0$

Two-sample *t*-test assumptions

Both samples must be distributed normally

Samples should have equal variances

- **F-test** can compare variances of samples to check assumption, but it is <u>highly sensitive and will "too often" reject the null</u>.
- **Levene's test** will test for homogeneity of variances as well
- Can use **Welch's *t*-test** when variance assumption is not met

For this class, we focus on normal assumption

Two-sample vs. paired t-test

Paired t-test is a special case where the two samples being compared have a *natural pairing*

- Effectively a **one-sample t-test** where we test the if *difference* between two samples = 0

Paired t-test must check assumption that *difference* between means is normal

You can always perform a two-sample instead of paired, but paired will have more power

Paired scenarios

Making two measurements on each subject

Making repeat measurements on the same subject at two time points

- Before and after treatment

Matching subjects with similar age, sex, etc.

Placing subject and control in close proximity

Is it paired or independent?

Triglyceride levels of a group of subjects is compared before and after taking a vitamin supplement. **Paired**

For a clinical trial, one group is given a vitamin and the other a placebo. Their triglyceride levels are compared. **Independent**

For a clinical trial, two groups with individuals matched for age, sex, and health history are given vitamin and placebo, respectively. Triglyceride levels are compared. **Paired**

I measure free energies, between inactive and active conformations, for 30 enzymes. **Paired**

A clinical trial tests a new drug on 20 sets of twins, giving, for each twin pair, one a placebo and one the drug. Comparisons between placebo and drug groups are made. **Paired**

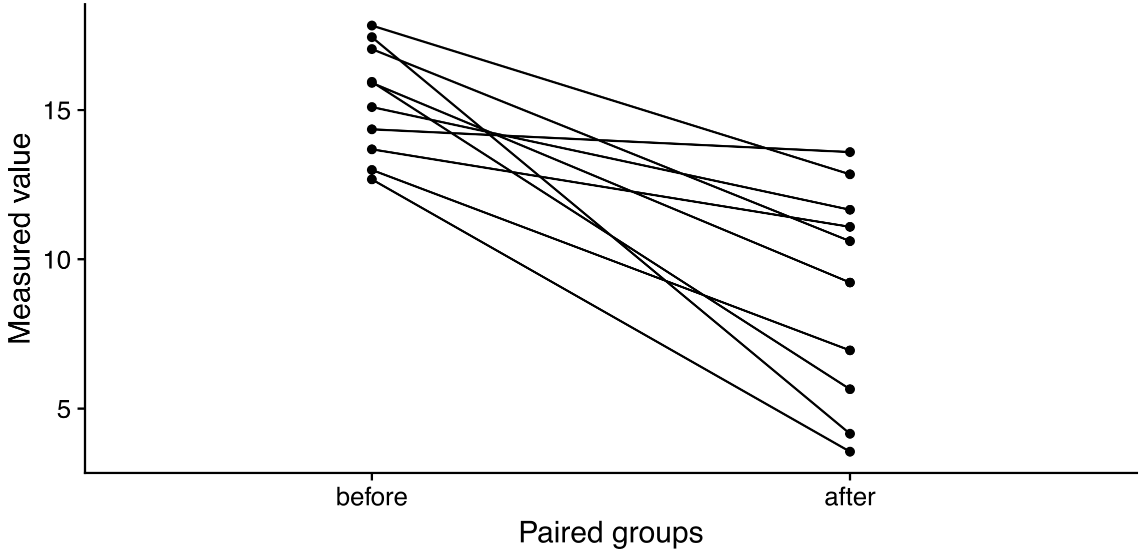
A clinical trial tests a new drug on 20 sets of twins, randomizing all individuals into a placebo group and or a drug group. Comparisons between groups are made. **Independent**

Is it paired or independent?

Paired: Can I draw a line between individuals in groups?

- Must be same N per group, by definition

Independent: No natural way to draw the lines.



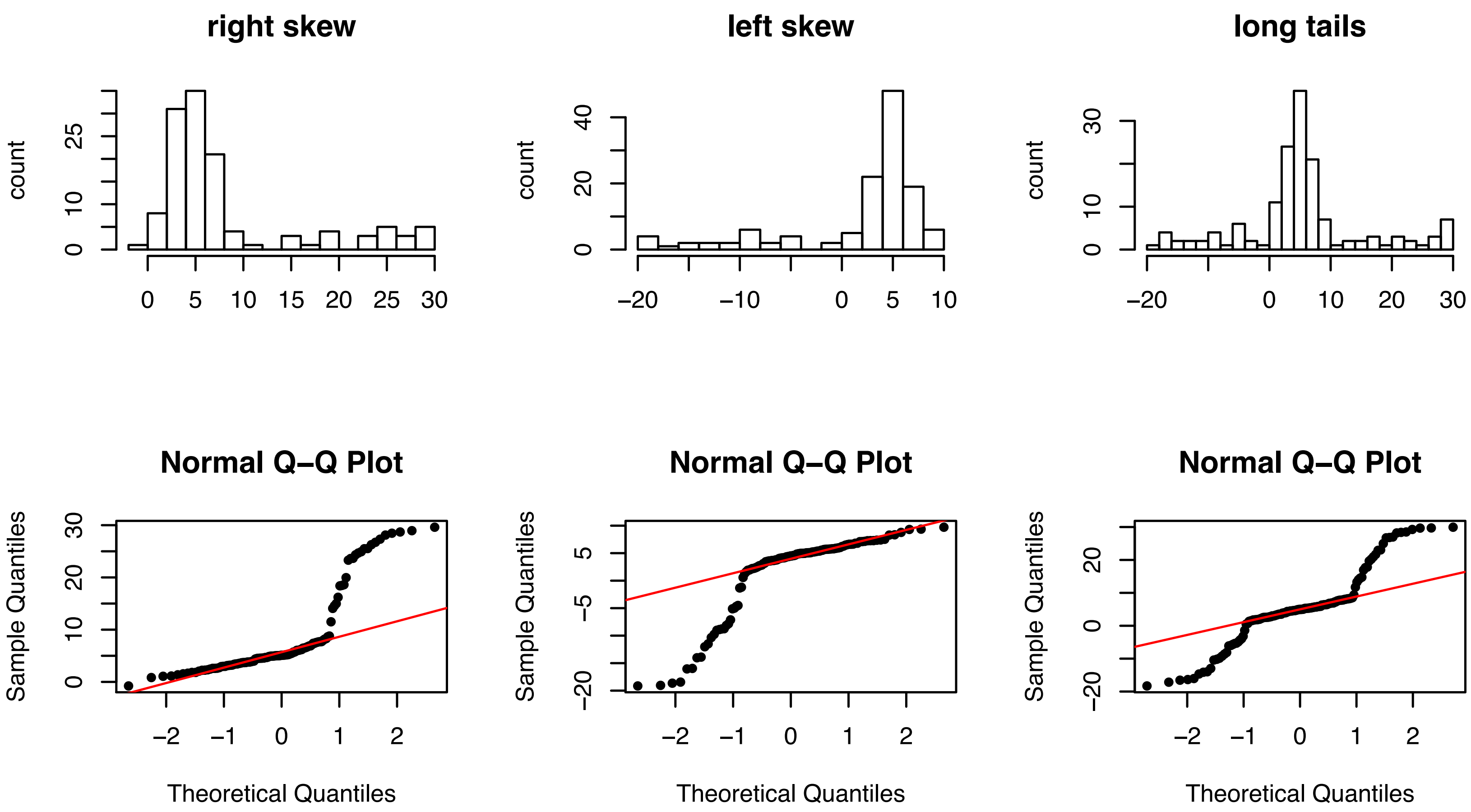
Troubleshooting: Failure to meet normal assumption

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: $x \rightarrow \log(x)$

Square root of data: $x \rightarrow \text{sqrt}(x)$

Inverse of data: $x \rightarrow 1/x$

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

Instructor Editorializing:

Much like Z-tables, data transforms are mostly historical