Comparing one mean to mean under null hypothesis

One-sample *t*-test

The one-sample t-test compares the mean of a random sample from a normal population with the population mean proposed in a null hypothesis.

Test statistic for one-sample *t*-test

$$t = \frac{\bar{Y} - \mu_0}{s/\sqrt{n}}$$

μ_0 is the mean value proposed by H_0

Hypotheses for one-sample *t*-tests

H_0 : *The mean of the population is* μ_0.

H_A: *The mean of the population is not* μ_0.

Example: Human body temperature

H_0 : Mean healthy human body temperature is 98.6ºF.

H_A: Mean healthy human body temperature is not 98.6ºF.

Human body temperature

$$n = 24$$

$$\bar{Y} = 98.28$$

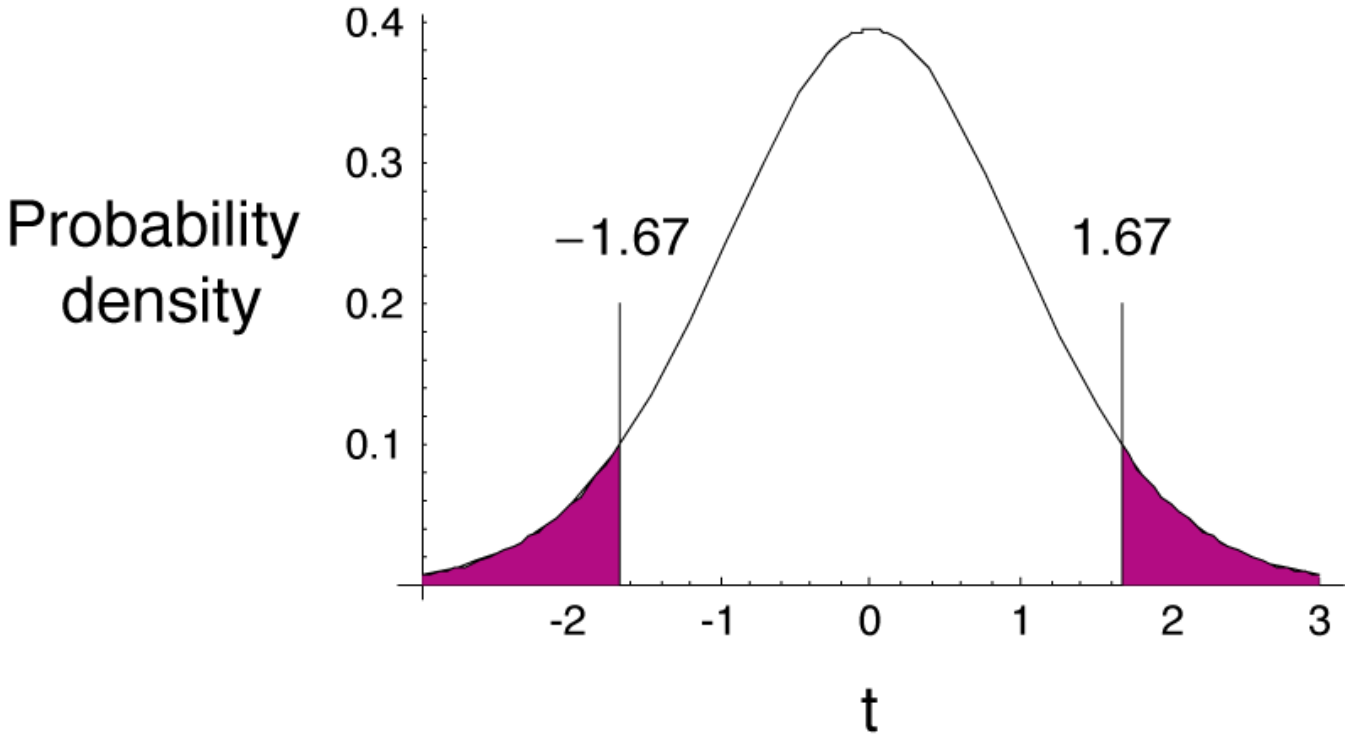
$$s = 0.940$$

$$t = \frac{\bar{Y} - \mu_0}{s/\sqrt{n}} = \frac{98.28 - 98.6}{0.940/\sqrt{24}} = -1.67$$

Degrees of freedom

$$df = n - 1 = 23$$

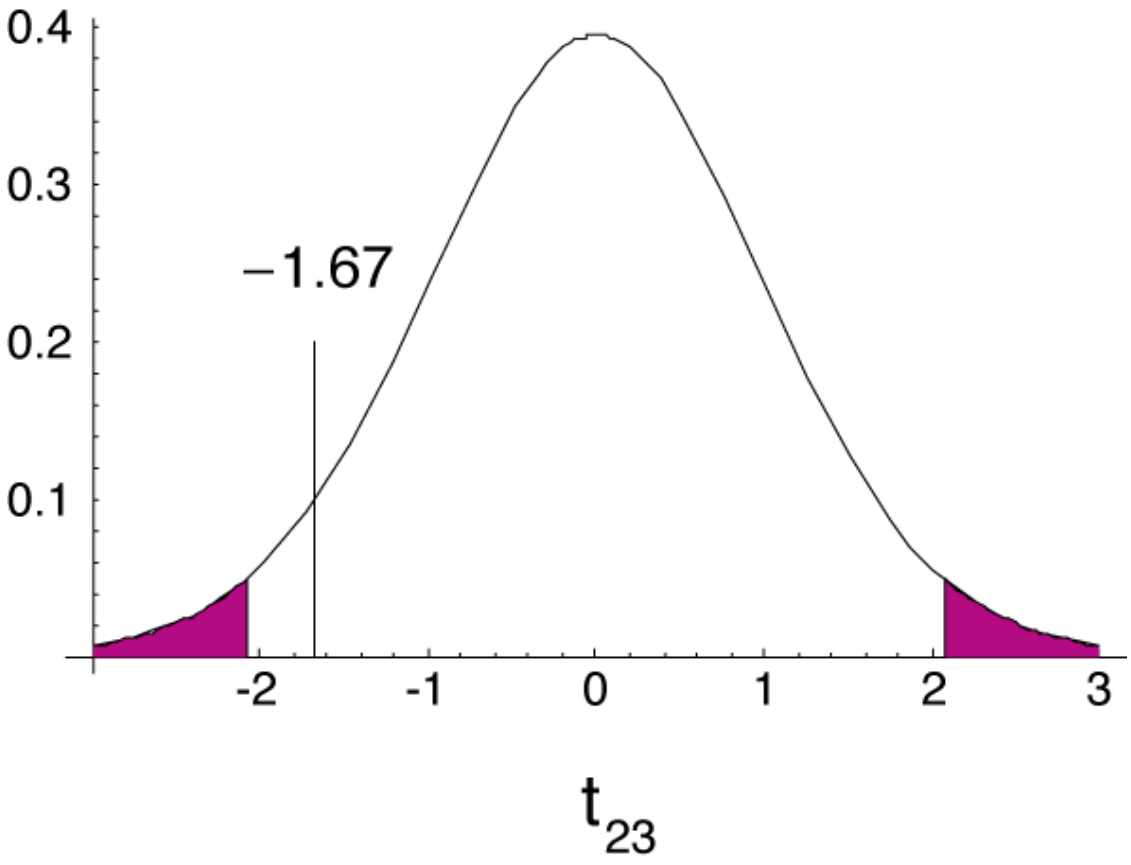
Comparing *t* to its distribution to find the *P*-value



A portion of the *t* table

df	α(1) =0.1 α(2)=0.2	α(1) =0.05 α(2)=0.10	α(1) =0.025 α(2)=0.05	α(1) =0.01 α(2)=0.02	α(1) =0.005 α(2)=0.01
...	...	...	...	...	...
20	1.33	1.72	2.09	2.53	2.85
21	1.32	1.72	2.08	2.52	2.83
22	1.32	1.72	2.07	2.51	2.82
23	1.32	1.71	2.07	2.5	2.81
24	1.32	1.71	2.06	2.49	2.8
25	1.32	1.71	2.06	2.49	2.79

qt(0.975, df = 23)


0.4
0.3
0.2
0.1
−1.67
-2
-1
0
1
2
3
t23

-1.67 is closer to 0 than -2.07, so P > 0.05.

With these data, we cannot reject the null hypothesis that the mean human body temperature is 98.6.

Body temperature revisited: $n = 130$

$$n = 130$$

$$\bar{Y} = 98.25$$

$$s = 0.733$$

$$t = \frac{\bar{Y} - \mu_0}{s/\sqrt{n}} = \frac{98.25 - 98.6}{0.733/\sqrt{130}} = -5.44$$

Body temperature revisited: $n = 130$

$$t = -5.44$$

$$t_{0.05(2),129} = \pm 1.98$$

t is further out in the tail than the critical value, so we could reject the null hypothesis. Human body temperature is not 98.6ºF.

One-sample *t*-test: Assumptions

- The variable is normally distributed.
- The sample is a random sample.

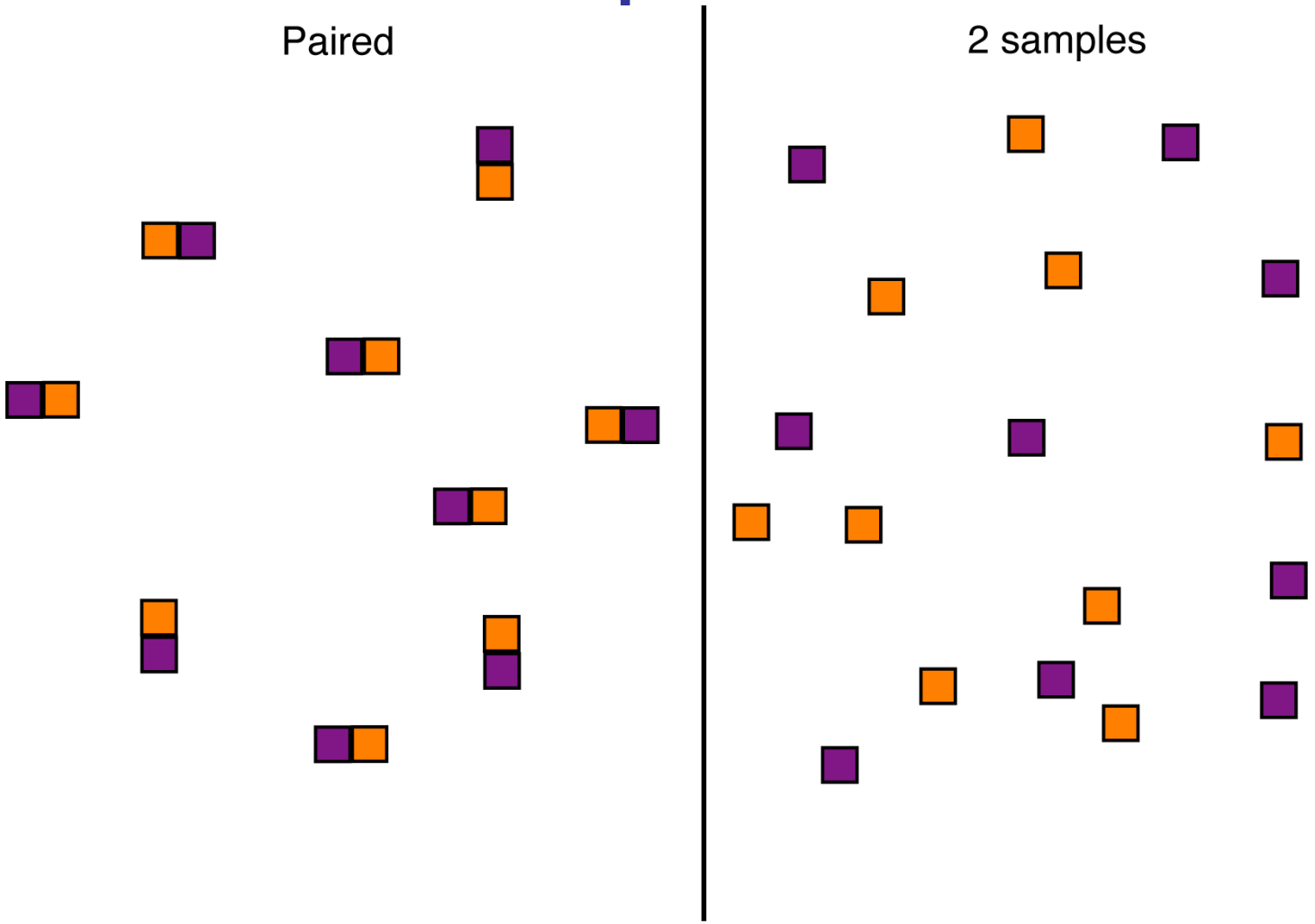
Comparing two means

- paired t-test
- comparison means of 2 samples
- two sample t-test
- differences in variance between two groups

Comparing means

- Situation: two variables, one categorical and one numerical
- Goal: to compare the mean of a numerical variable for different groups.

Paired vs. 2 sample comparisons



Paired comparisons allow us to account for a lot of extraneous variation

2 sample methods are sometimes easier in terms of data collection

Paired designs

- Data from the two groups are paired
- Each member of the pair shares much in common with the other, *except* for the tested categorical variable
- There is a one-to-one correspondence between the individuals in the two groups

Paired design: Examples

- Patients before vs. after treatment
- Rivers upstream vs. downstream of power plants
- Identical twins: one with a treatment and one without
- Earwigs in each ear: how to get them out? Compare tweezers to hot oil

Paired comparisons

- We have many pairs
- In each pair, there is one member that has one treatment and another who has another treatment

  (“Treatment” can mean “group”)

Paired comparisons

- To compare two groups, we use the mean of the *difference* between the two members of each pair

Paired *t* test

- Compares the mean of the differences to a value given in the null hypothesis

- For each pair, calculate the difference. The paired *t*-test is simply a one-sample *t*-test on the differences.

Example: National No Smoking Day

- Data compares injuries at work on National No Smoking Day (in Britain) to the same day the week before

- Each data point is a year

Data from Waters et al. (1998) Nicotine withdrawal and accident rates. *Nature* 394: 137.

Year	Injuries before No Smoking Day	Injuries on No Smoking Day
1987	516	540
1988	610	620
1989	581	599
1990	586	639
1991	554	607
1992	632	603
1993	479	519
1994	583	560
1995	445	515
1996	522	556

Data from Waters et al. (1998) Nicotine withdrawal and accident rates. *Nature* 394: 137.

Hypotheses

H_0: The number of work related injuries **is not different** between No Smoking Days and other comparable days. ($\mu_d = 0$)

H_A: The number of work related injuries **is different** between No Smoking Days and other comparable days. ($\mu_d \neq 0$)

Calculate differences

Injuries before No Smoking Day	Injuries on No Smoking Day	Difference (d)
516	540	24
610	620	10
581	599	18
586	639	53
554	607	53
632	603	-29
479	519	40
583	560	-23
445	515	70
522	556	34

Calculate *t* using *d*'s

$$\bar{d} = 25$$

$$s_d^2 = 1043.78$$

$$n = 10$$

$$t = \frac{\bar{d} - \mu_0}{SE_{\bar{d}}} = \frac{25 - 0}{\sqrt{1043.78/10}} = 2.45$$

CAUTION!

- The number of data points in a paired *t* test is the number of *pairs*. -- ***Not*** the number of individuals

- Degrees of freedom = Number of pairs - 1

Critical value of *t*

$$t_{0.05(2),9} = 2.26$$

$$t_{data} = 2.45 > 2.26$$

So we can reject the null hypothesis of no relationship.

Our data indicate that stopping smoking increases job-related accidents in the short term.

Assumptions of paired *t* test

- Pairs are chosen at random
- The differences have a normal distribution

It does *not* assume that the individual values are normally distributed, only that the differences are.

Comparing the means of two groups (not paired)

1) Estimate the difference in population means, and your uncertainty on that estimate: *Confidence Interval*

2) Hypothesis test: *2-sample t test*

Estimation: Difference between two means

$$\bar{Y}_1 - \bar{Y}_2$$

Confidence interval: $\left(\bar{Y}_1 - \bar{Y}_2\right) \pm SE_{\bar{Y}_1 - \bar{Y}_2}\, t_{\alpha(2),df}$

SE of difference means

Standard error of difference between means

$$SE_{\bar{Y}_1 - \bar{Y}_2} = \sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}$$

pooled variance

Standard error of difference between means

$$SE_{\bar{Y}_1 - \bar{Y}_2} = \sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}$$

Pooled variance: $$s_p^2 = \frac{df_1 s_1^2 + df_2 s_2^2}{df_1 + df_2}$$

$df_1 = n_1 - 1$; $df_2 = n_2 - 1$

Costs of resistance to aphids

2 genotypes of lettuce: *Susceptible* and *Resistant*

Do these genotypes differ in fitness in the absence of aphids?

Data, summarized

	Susceptible	Resistant
Mean number of buds	720	582
SD of number of buds	223.6	277.3
Sample size	15	16

Both distributions are approximately normal.

The confidence interval of the difference in the means

$$\left(\bar{Y}_1 - \bar{Y}_2\right) \pm SE_{\bar{Y}_1 - \bar{Y}_2}\, t_{\alpha(2),df}$$

Calculating the standard error

$df_1 = 15 - 1 = 14; \quad df_2 = 16 - 1 = 15$

$$s_p^2 = \frac{df_1 s_1^2 + df_2 s_2^2}{df_1 + df_2} = \frac{14(223.6)^2 + 15(277.3)^2}{14 + 15} = 63909.9$$

Calculating the standard error

$df_1 = 15 - 1 = 14$; $df_2 = 16 - 1 = 15$

$$s_p^2 = \frac{df_1 s_1^2 + df_2 s_2^2}{df_1 + df_2} = \frac{14(223.6)^2 + 15(277.3)^2}{14 + 15} = 63909.9$$

$$SE_{\bar{Y}_1 - \bar{Y}_2} = \sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}} = \sqrt{\frac{63909.9}{15} + \frac{63909.9}{16}} = 90.86$$

Finding critical value of *t*
(for 95% C.I.)

$$df = df_1 + df_2 = n_1 + n_2 - 2$$

$$= 15 + 16 - 2$$
$$= 29$$

$$t_{0.05(2),29} = 2.05$$

The 95% confidence interval of the difference in the means

$$\left(\bar{Y}_1 - \bar{Y}_2\right) \pm SE_{\bar{Y}_1 - \bar{Y}_2}\, t_{\alpha(2),df} = \left(720 - 582\right) \pm 90.86(2.05)$$

$$= 138 \pm 186 \text{ buds per plant}$$

Notice that the confidence interval on the difference in means includes zero. This is telling us that, given the data, no difference in buds is plausible.

Testing hypotheses about the difference in two means

2-sample *t*-test

The *two sample* t-*test* compares the means of a numerical variable between two populations.

Test statistic in 2-sample t-test

$$t = \frac{\bar{Y}_1 - \bar{Y}_2}{SE_{\bar{Y}_1 - \bar{Y}_2}}$$

Hypotheses

H_0: There is no difference between the number of buds in the susceptible and resistant plants. ($\mu_1 = \mu_2$)

H_A: The resistant and the susceptible plants differ in their mean number of buds. ($\mu_1 \neq \mu_2$)

Calculating *t*

$$t = \frac{\left(\bar{Y}_1 - \bar{Y}_2\right)}{SE_{\bar{Y}_1 - \bar{Y}_2}} = \frac{\left(720 - 582\right)}{90.86} = 1.52$$

Drawing conclusions...

$$df = n_1 + n_2 - 2 = 29$$

$$t_{0.05(2),29} = 2.05$$

$t = 1.52 < 2.05$, so we cannot reject the null hypothesis.

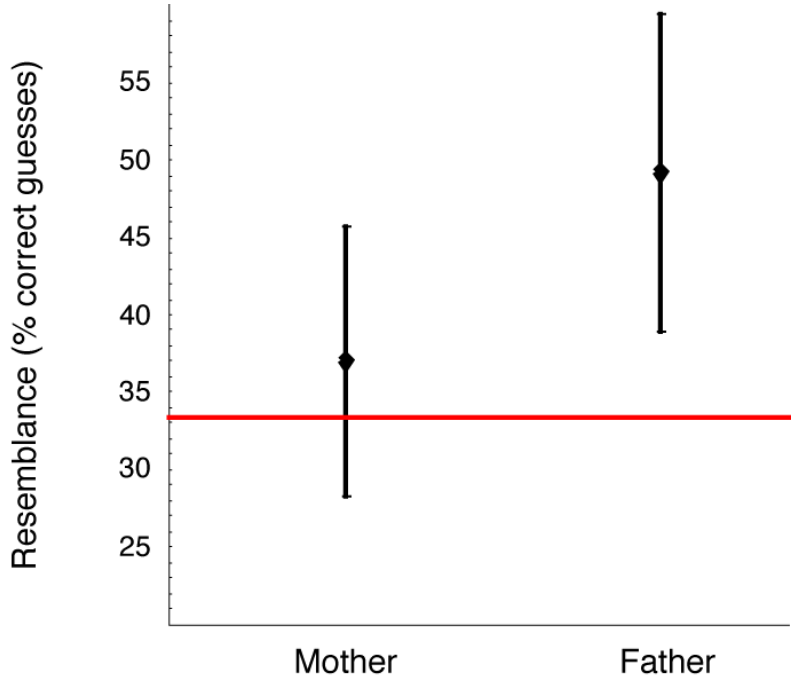
These data are not sufficient to say that there is a cost of resistance.

Assumptions of two-sample *t* -tests

- Both samples are random samples.
- Both populations have normal distributions
- The variance of both populations is equal.

The wrong way to make a comparison of two groups

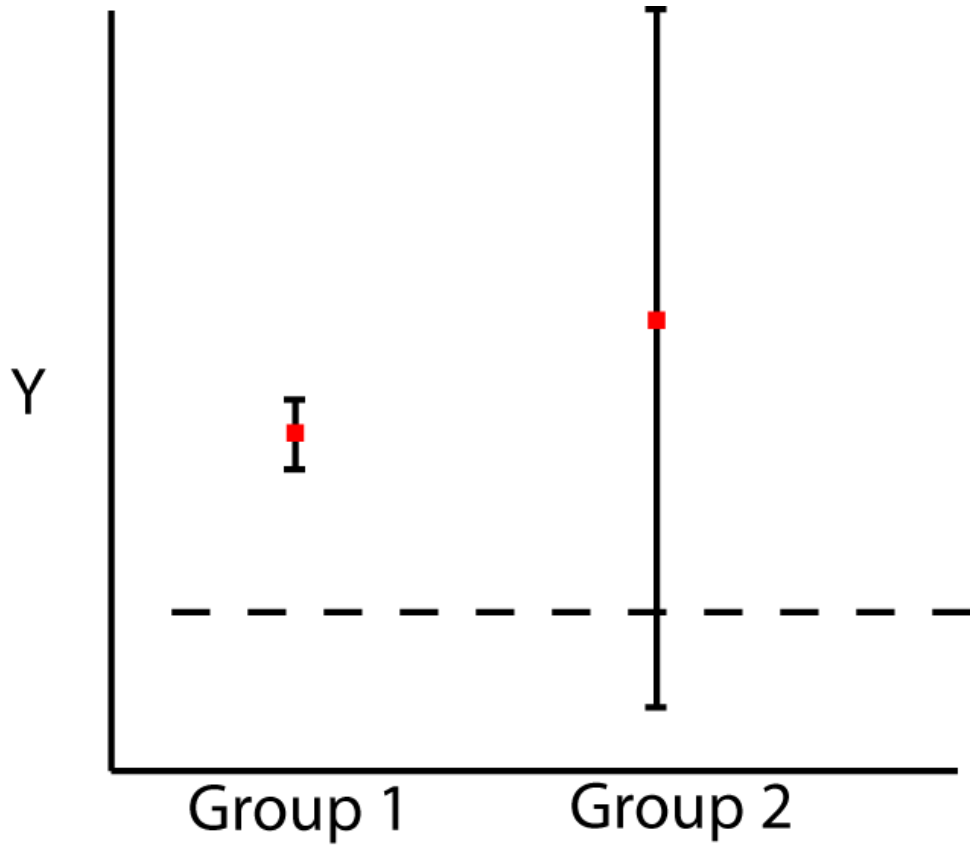
~~"Group 1 is significantly different from a constant, but Group 2 is not. Therefore Group 1 and Group 2 are different from each other."~~



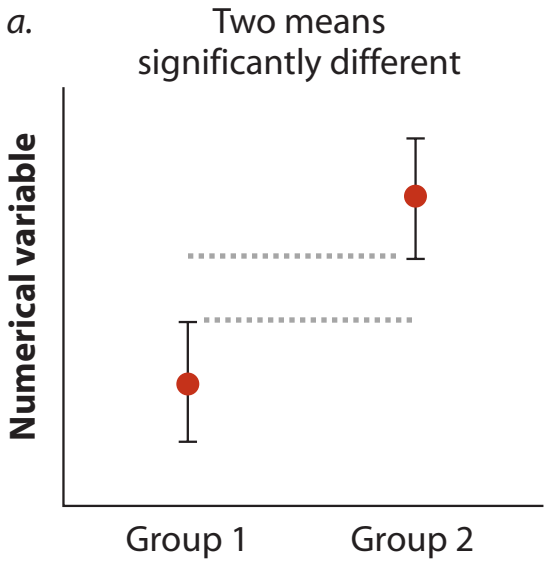
Error bars depict 95% Confidence Intervals

Christenfeld & Hill (1995) *Nature.*

A more extreme case...



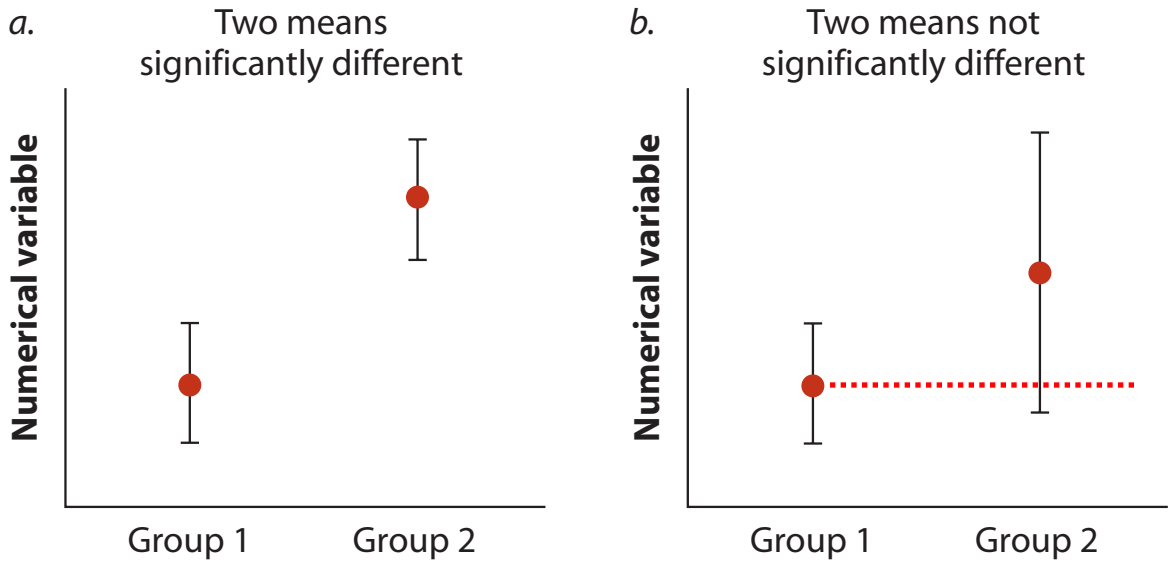
Interpreting overlap of CI's



no overlap CI's

H_0 *reject*

Interpreting overlap of CI's



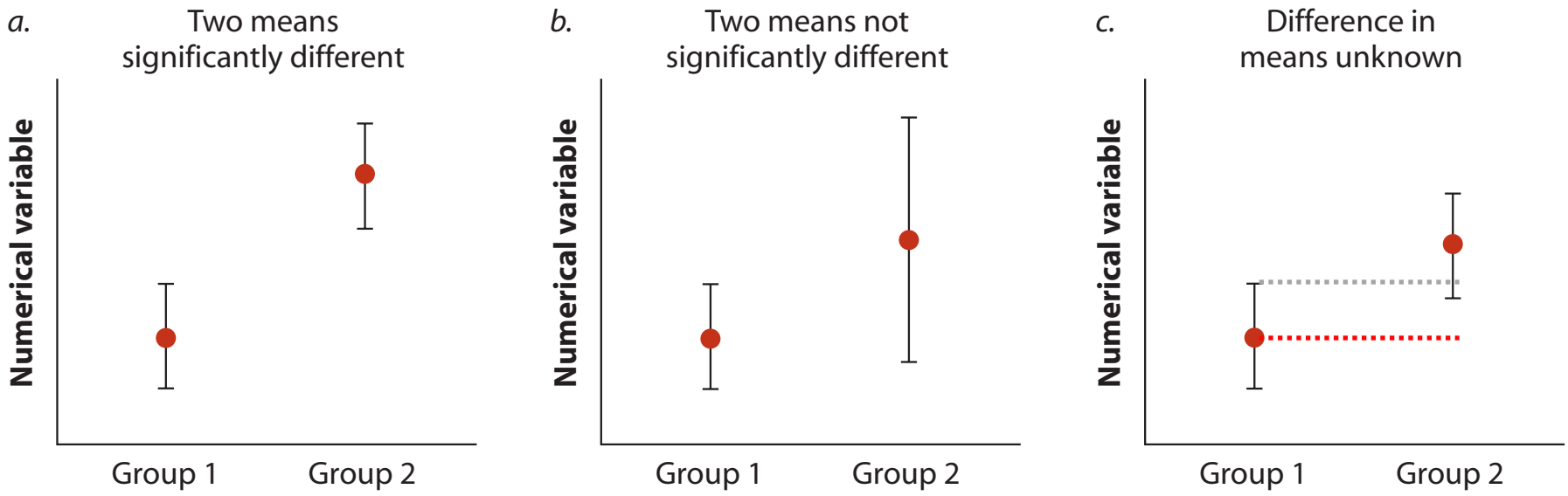
no overlap CI's

H_0 *reject*

CI overlaps mean

not reject

Interpreting overlap of CI's



	no overlap CI's	*CI overlaps mean*	*partial overlap*
H_0	*reject*	*not reject*	*unknown*

Comparing means when variances are not equal

Welch's *t* test

Welch's approximate t-test compares the means of two normally distributed populations that have unequal variances.

Burrowing owls and dung traps

Levey et al. (2004) Use of dung as a tool by burrowing owls. *Nature* 431: 39

Dung beetles

Experimental design

- 20 randomly chosen burrowing owl nests
- Randomly divided into two groups of 10 nests
- One group was given extra dung; the other not
- Measured the number of dung beetles in the owls' diets

Number of beetles caught

- Dung added: $\bar{Y} = 4.8$

  $s = 3.26$

- No dung added: $\bar{Y} = 0.51$

  $s = 0.89$

Hypotheses

H_0: Owls catch the same number of dung beetles with or without extra dung ($\mu_1 = \mu_2$).

H_A: Owls do not catch the same number of dung beetles with or without extra dung ($\mu_1 \neq \mu_2$).

Welch's *t*

$$t = \frac{\bar{Y}_1 - \bar{Y}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

$$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\left(\frac{\left(s_1^2/n_1\right)^2}{n_1 - 1} + \frac{\left(s_2^2/n_2\right)^2}{n_2 - 1}\right)}$$

Round down *df* to nearest integer

Owls and dung beetles

$$t = \frac{\bar{Y}_1 - \bar{Y}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{4.8 - 0.51}{\sqrt{\frac{3.26^2}{10} + \frac{0.89^2}{10}}} = 4.01$$

Degrees of freedom

$$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\left(\frac{\left(s_1^2/n_1\right)^2}{n_1 - 1} + \frac{\left(s_2^2/n_2\right)^2}{n_2 - 1}\right)} = \frac{\left(\frac{3.26^2}{10} + \frac{0.89^2}{10}\right)^2}{\left(\frac{\left(3.26^2/10\right)^2}{10 - 1} + \frac{\left(0.89^2/10\right)^2}{10 - 1}\right)} = 10.33$$

Which we round down to *df* = 10

(In a regular *t*-test, the *df* would be 18)

Reaching a conclusion

$$t_{0.05(2),10} = 2.23$$

$$t = 4.01 > 2.23$$

So we can reject the null hypothesis with $P < 0.05$.

Extra dung near burrowing owl nests increases the number of dung beetles eaten.

Comparing the variance of two groups

$$H_0 : \sigma_1^2 = \sigma_2^2$$

$$H_A : \sigma_1^2 \neq \sigma_2^2$$

One possible method: the *F* test

The test statistic *F*

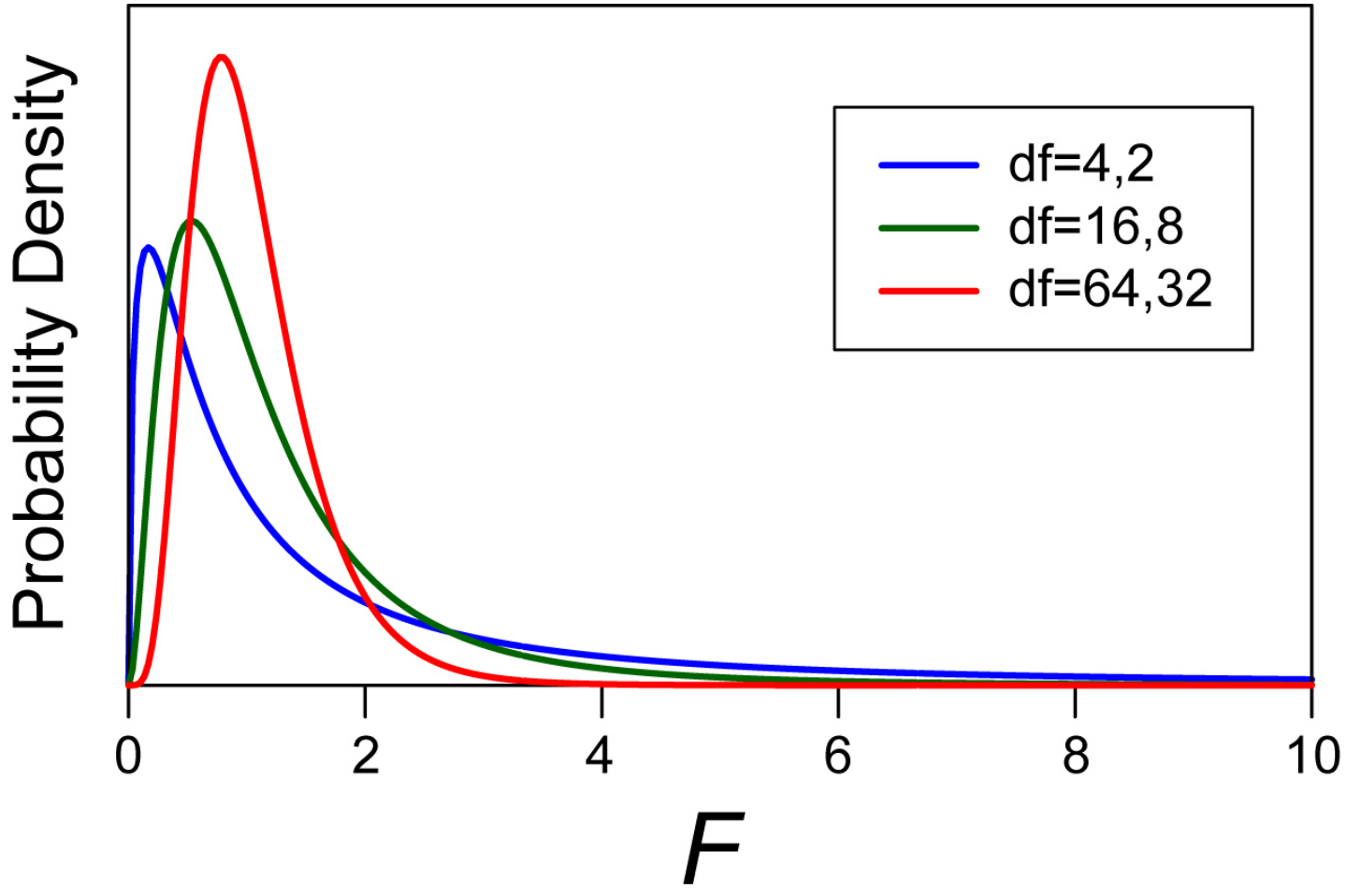
$$F = \frac{s_1^2}{s_2^2}$$

Put the larger s^2 on top in the numerator.

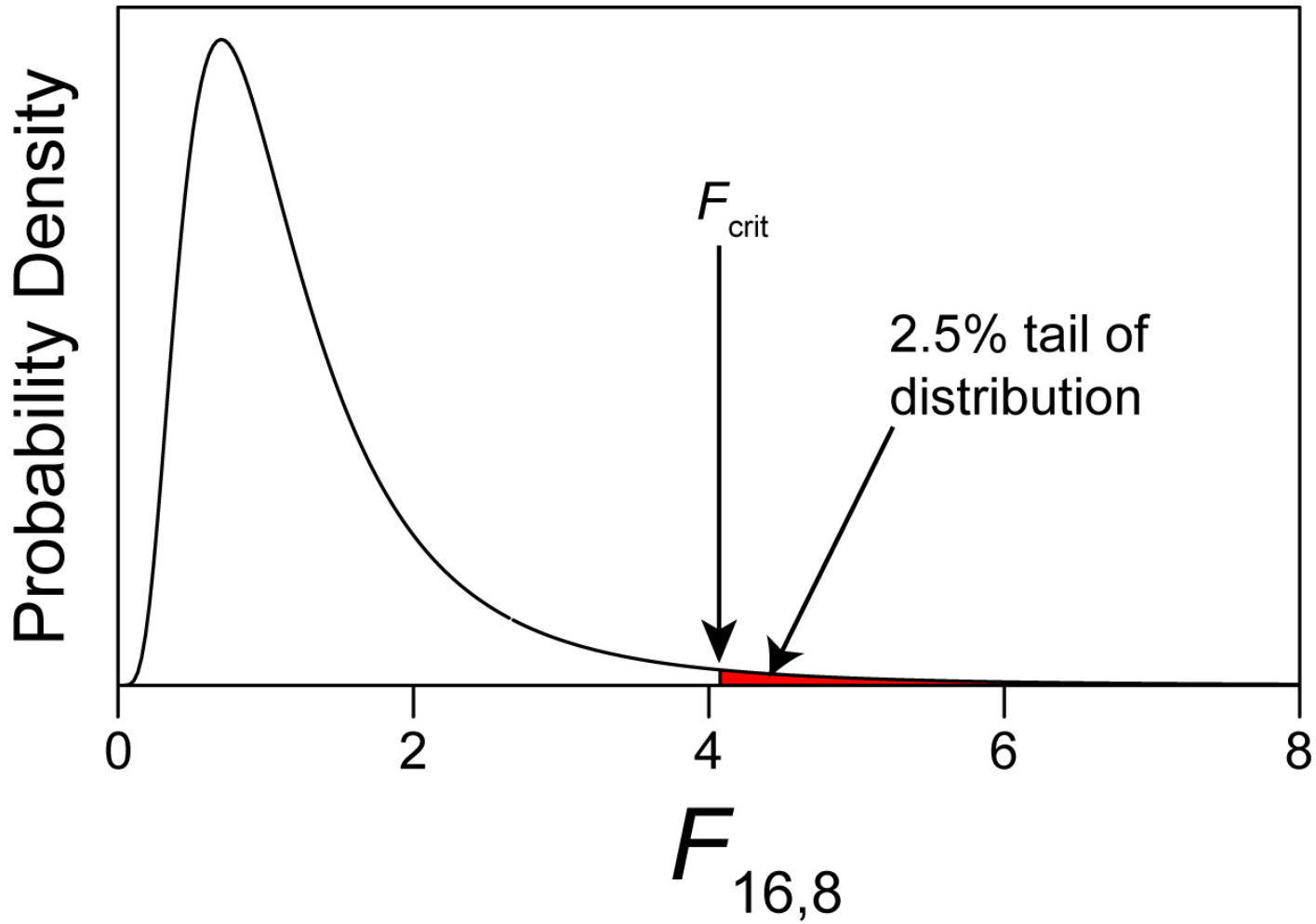
F...

- *F* has two different degrees of freedom, one for the numerator and one for the denominator. (Both are $df = n_i - 1$.) The numerator *df* is listed first, then the denominator *df*.

Example F distributions

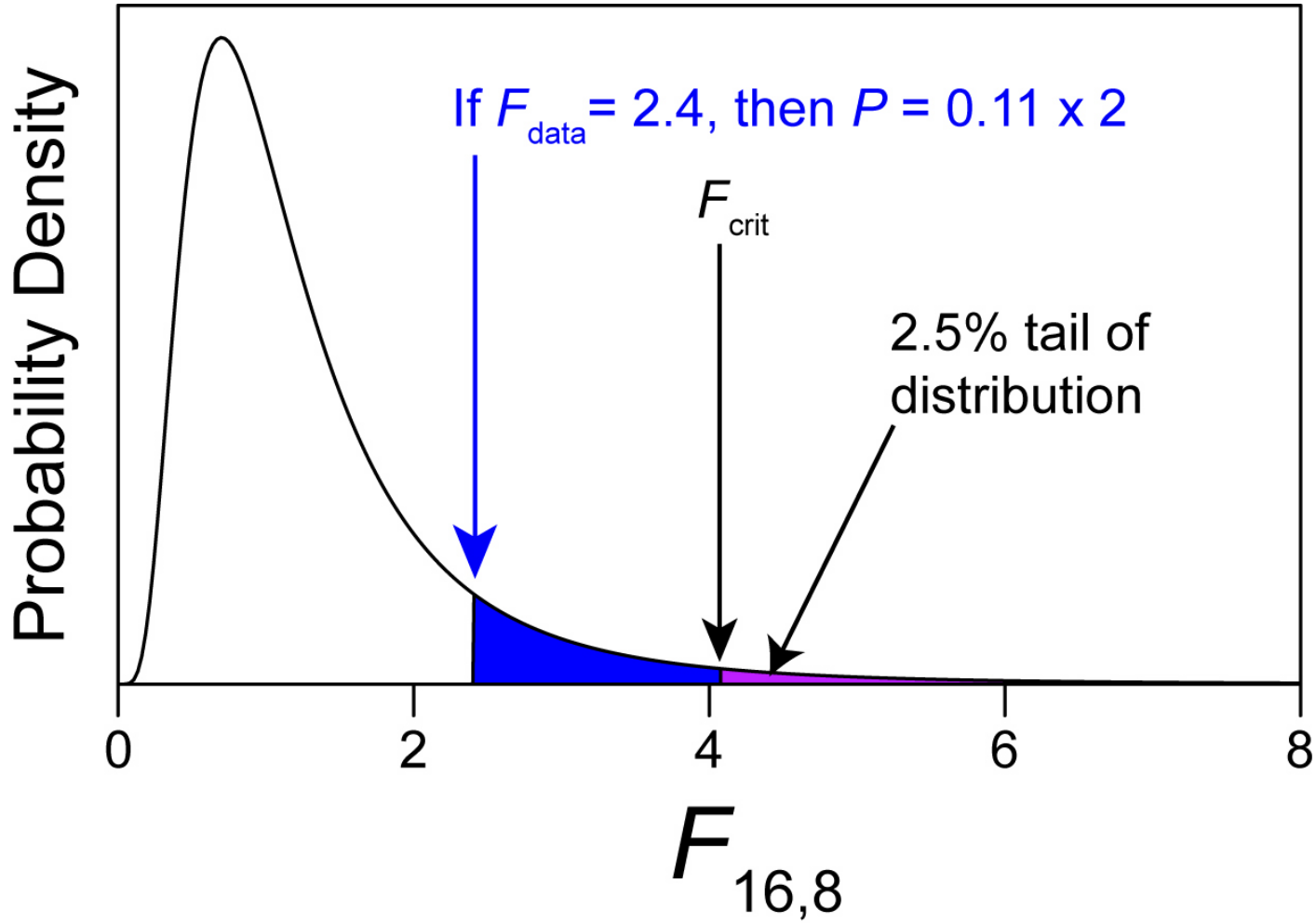


F_{crit} is the value of F where 2.5% of the distribution is to the right



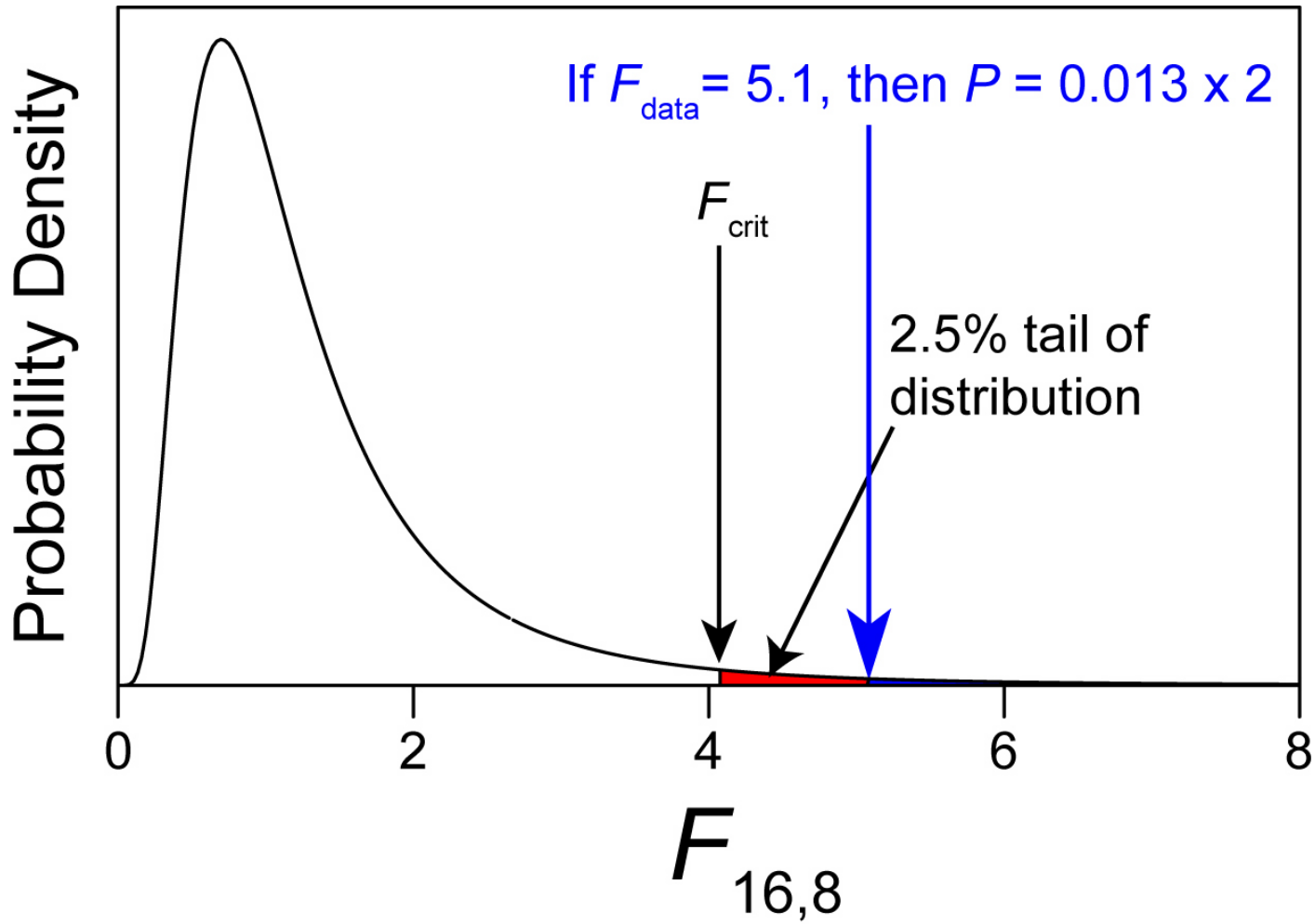
F_{data} allows us to determine how strange the data are, assuming the null is true.

(P is the probability of getting an F that big or bigger, if the null is true)

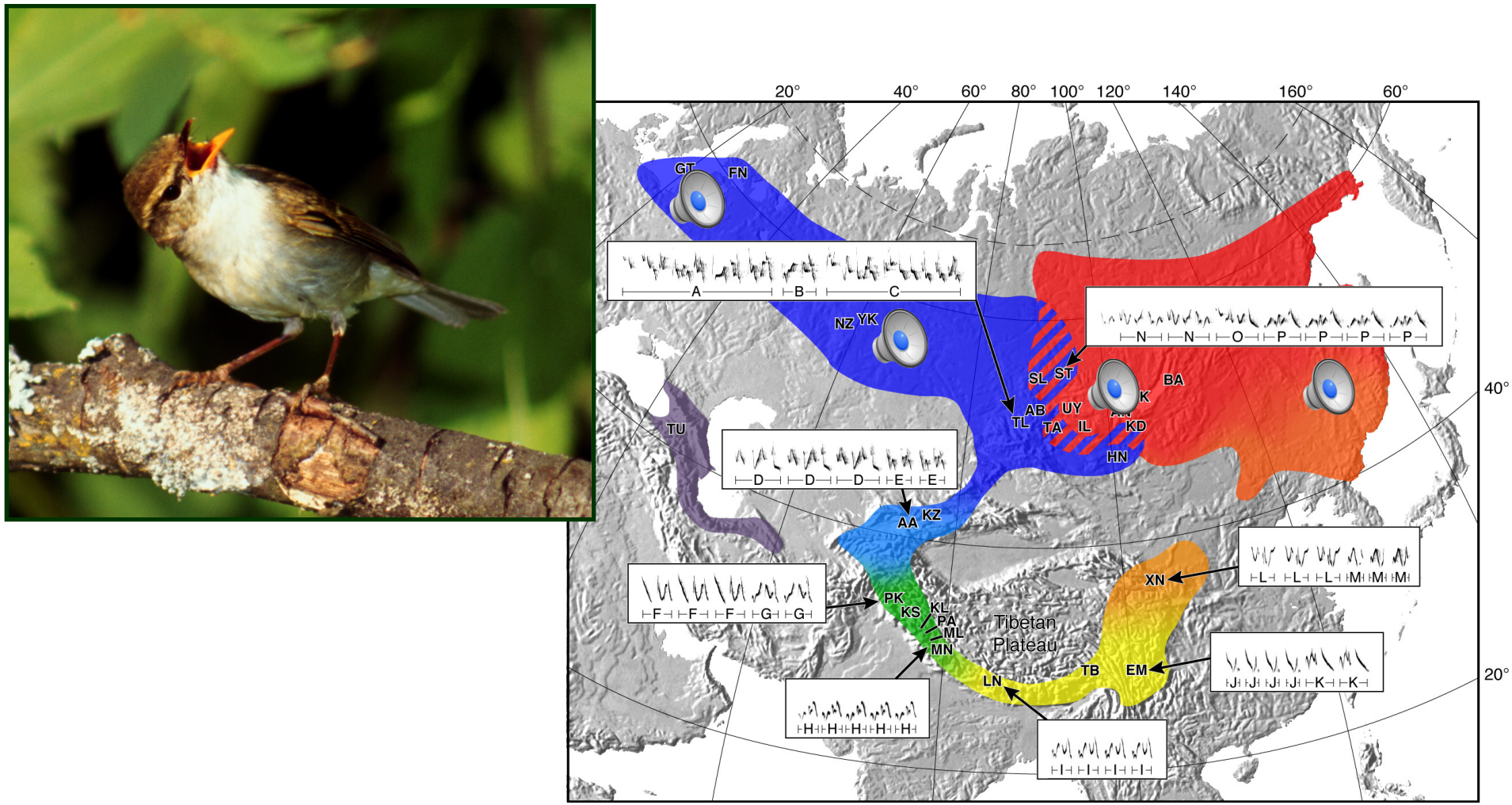


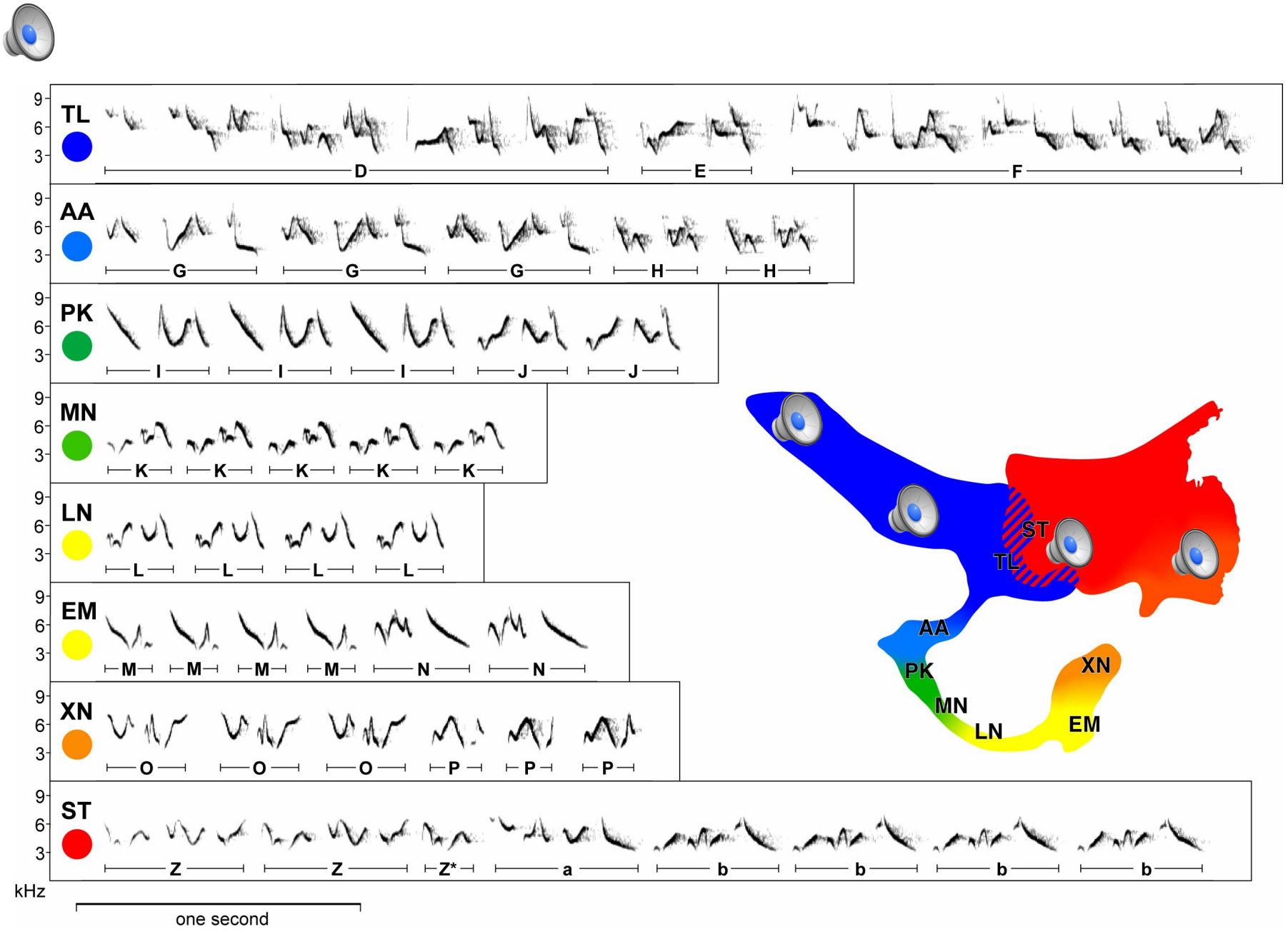
F_{data} allows us to determine how strange the data are, assuming the null is true.

(P is the probability of getting an F that big or bigger, if the null is true)



Example: Song length of greenish warblers




TL
9
6
3
D
E
F
AA
G
G
G
H
H
PK
I
I
I
J
J
MN
K
K
K
K
K
LN
L
L
L
L
EM
M
M
M
M
N
N
XN
O
O
O
P
P
P
ST
Z
Z
Z*
a
b
b
b
b
kHz
one second
ST
TL
AA
PK
MN
LN
XN
EM

Song lengths from India and Kyrgyzstan

India

Individual	Song length (s)
A	1.92
B	1.34
C	1.39
D	1.70
E	1.64

Mean: 1.60
Variance: 0.056
n: 5

Kyrgyzstan

Individual	Song length (s)
A	2.27
B	3.46
C	2.50
D	1.71

Mean: 2.48
Variance: 0.532
n: 4

Is there a difference in variation between sites?

$$s^2_{India} = 0.056 \qquad s^2_{Kyrgyz} = 0.532$$

$$F = \frac{0.532}{0.056} = 9.56$$

Determining the critical value of F

$$df_{Kyrgyz} = 4 - 1 = 3$$

$$df_{India} = 5 - 1 = 4$$

$$F_{crit} = F_{0.025,3,4} = 9.98$$

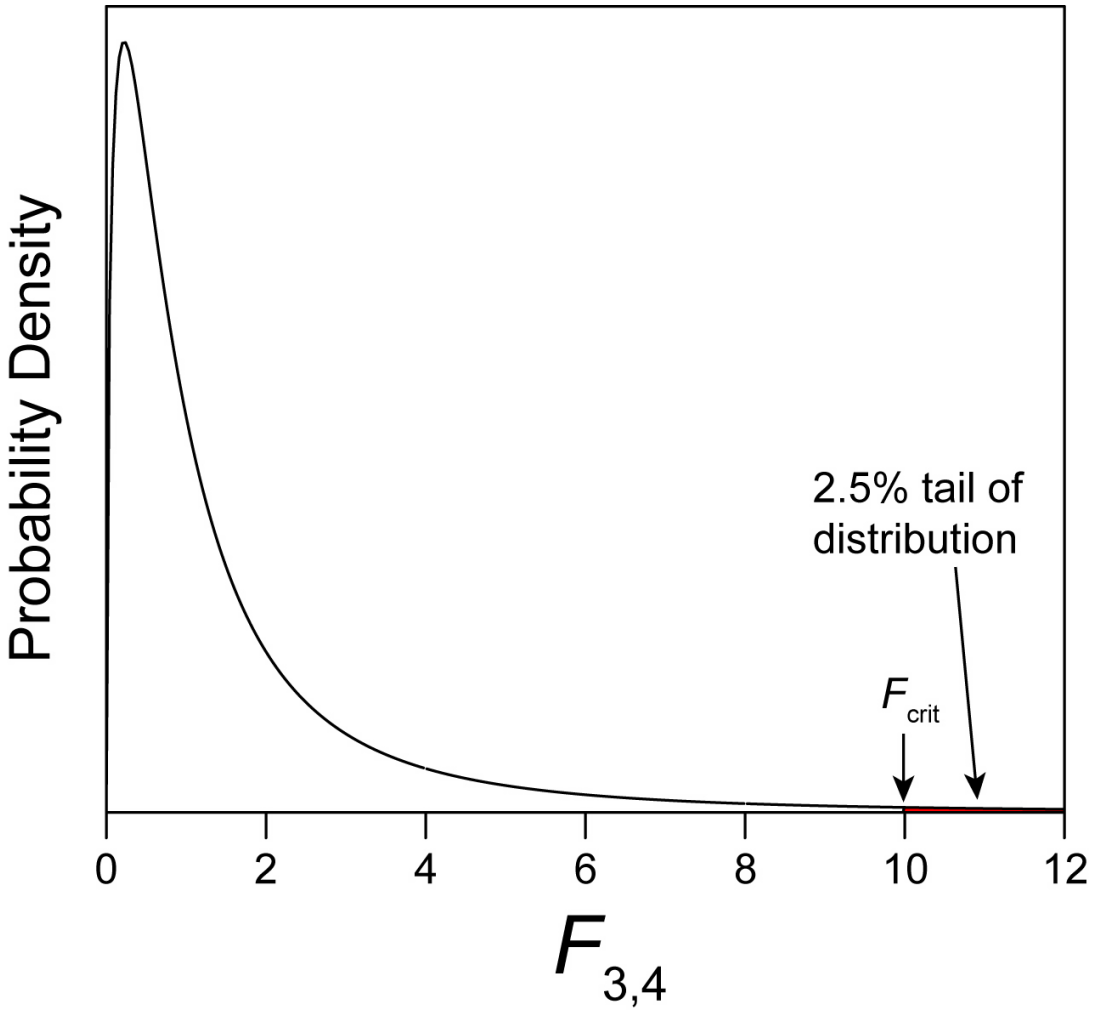
For a 2-tailed test, we compare to $F_{\alpha/2,df1,df2}$ from Table D.

Why $\alpha/2$ for the critical value?

By putting the larger s^2 in the numerator, we are forcing F to be greater than 1.

By the null hypothesis there is a 50:50 chance of either s^2 being greater, so we want the higher tail to include just $\alpha/2$.

F_{crit} is the value of F where 2.5% of the distribution is to the right



Conclusion

The $F = 9.56$ from the data is less than than $F_{(0.025),3,4} = 9.98$, so we cannot reject the null hypothesis that the variances of the two groups are equal.

The variance of song length might be larger in Kyrgyzstan than in India, but our data do not allow us to reject (at a confidence level of 95%) the null of no difference.

The *F* test is very sensitive to its assumption that both distributions are normal.

A more robust test to compare variances (between 2 or more groups) is:

Levene's test

You **should know** that Levene's test exists and why you would use it, but you do not need to know how to do it in this class. You would use a computer to do it, as the calculations are cumbersome.

Two-sample t-test not robust to unequal variances when:

- There is more than a 3-fold difference in standard deviation
- If the sample sizes of the two groups are very different or less than 30 with some difference in standard deviations

Wrap up

Paired vs. 2- sample

Paired → *take difference per pair* → *one sample t-test*

2-sample → *get confidence interval*

→ *2-sample t-test*

↓ *variance not equal*

Welch's t

Test variance → *F-test*

↓ *distributions not normal*

Levene's test

Handling violations of assumptions

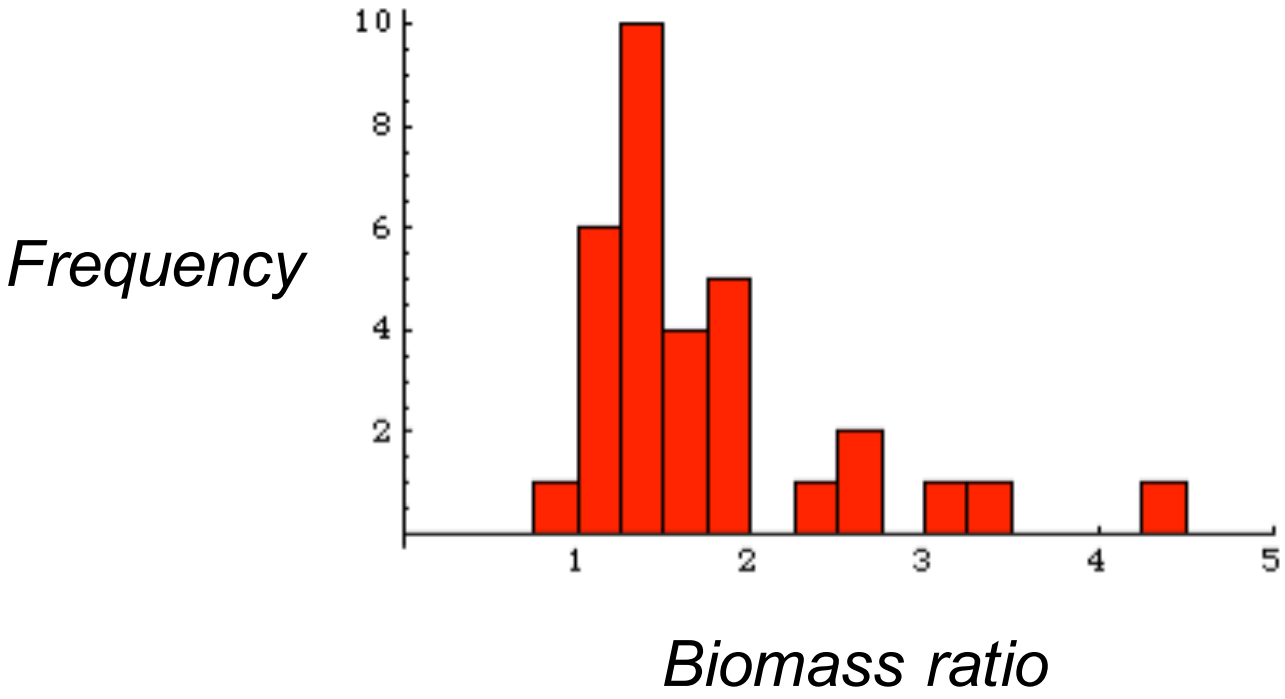
Assumptions of *t*-tests

- Random sample(s)
- Populations are normally distributed
- (for 2-sample *t*) Populations have equal variances

Detecting deviations from normality

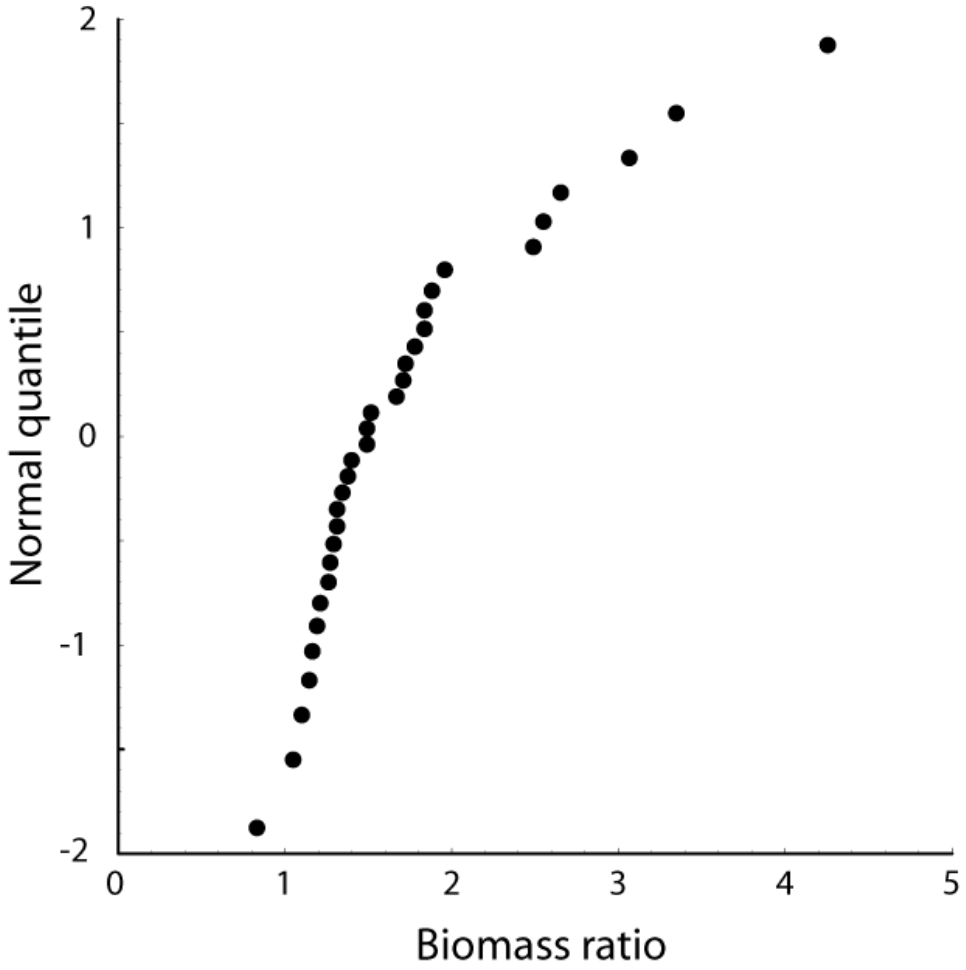
- Previous data/ theory
- Histograms
- Quantile plots
- Shapiro-Wilk test

Detecting deviations from normality: by histogram



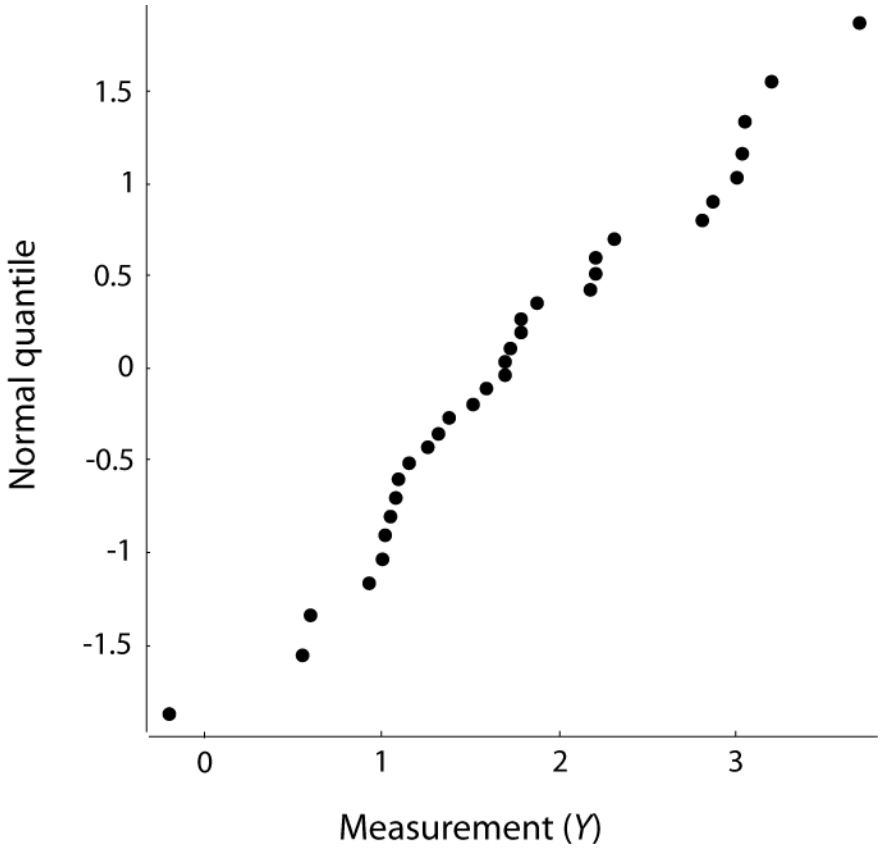
Halpern, B. S. 2003. The impact of marine reserves: Do reserves work and does reserve size matter? Ecological Applications 13:S117-S137.

Detecting deviations from normality: by quantile plot



Detecting deviations from normality: by quantile plot

Normal data



Detecting differences from normality: Shapiro-Wilk test

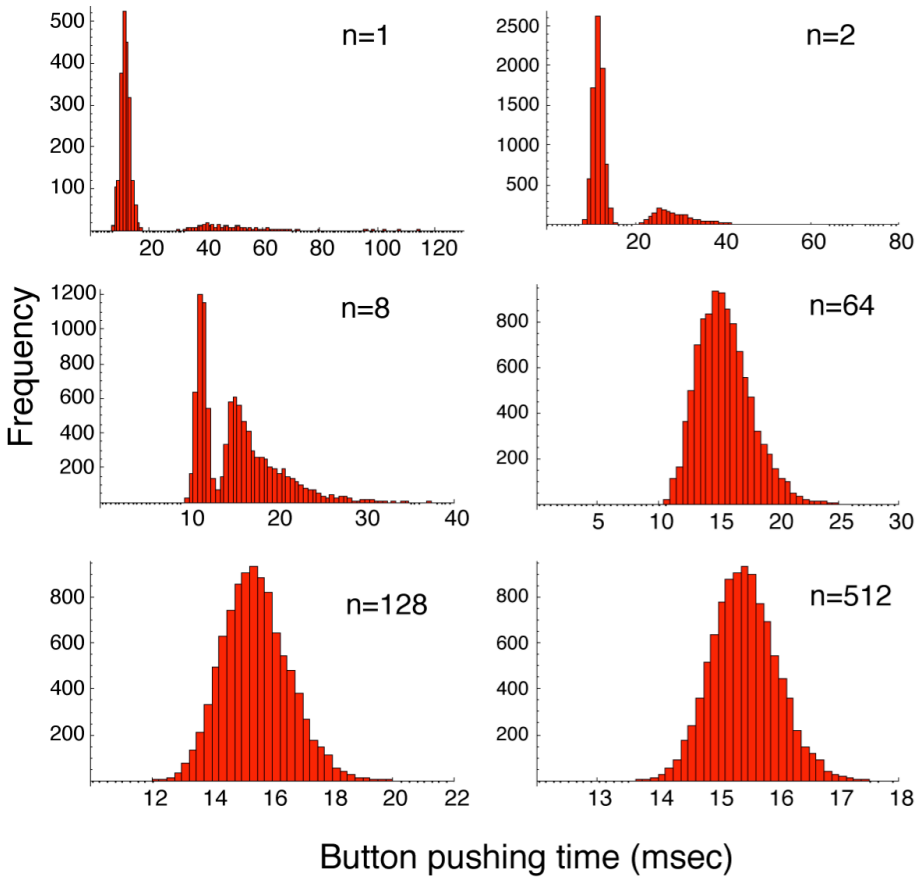
A Shapiro-Wilk test is used to test statistically whether a set of data comes from a normal distribution.

What to do when the assumptions are not true

- If the sample sizes are large, and deviation from normality not too big, parametric tests may be OK anyway
- Transformations
- Non-parametric tests
- Randomization and resampling

Central limit theorem

The sum or mean of a large number of measurements randomly sampled from any population is approximately normally distributed.

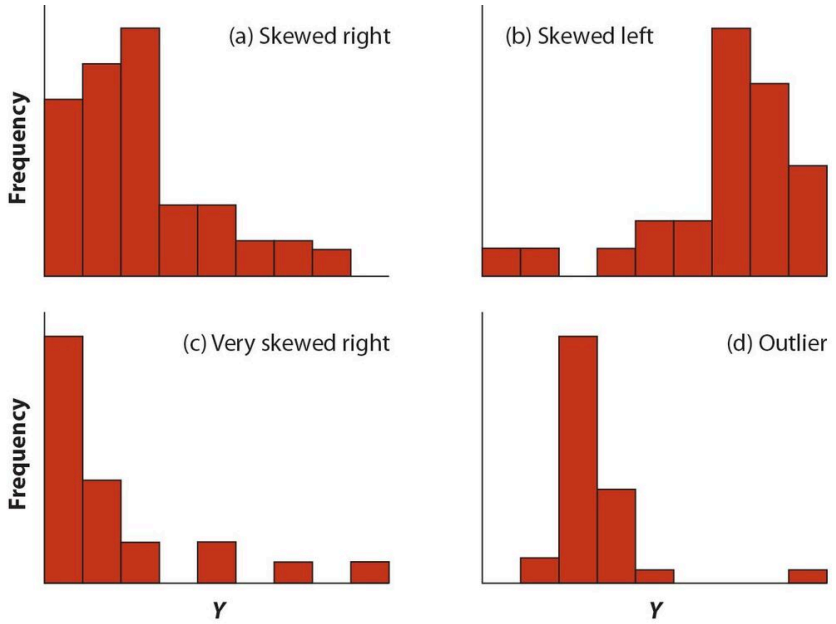


The normal approximation

- Means of large samples are normally distributed (roughly; this is the Central Limit Theorem)
- So, the parametric tests on large samples work relatively well, even for non-normal data.
- Rule of thumb- if n > ~50, the normal approximations may work

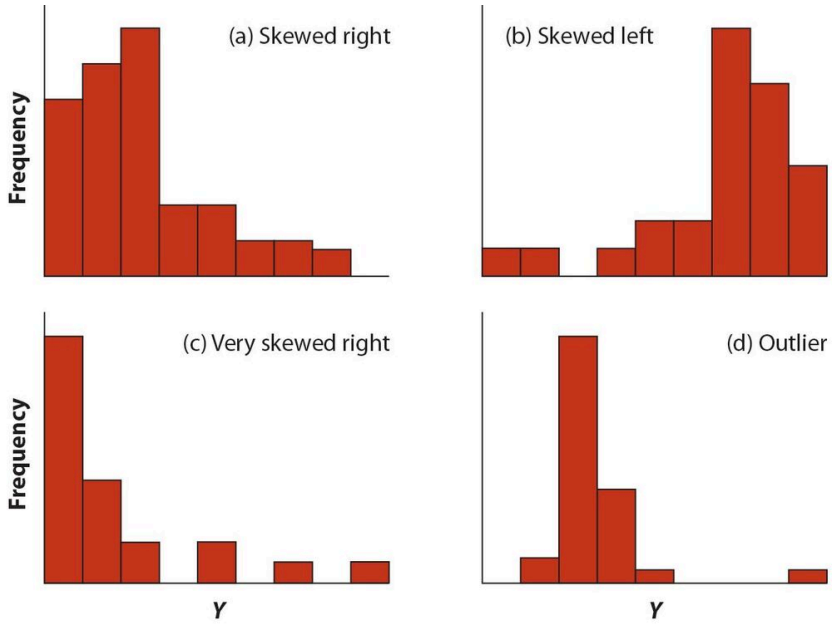
Take care

- Data which deviate from normality in different ways (e.g. a and b) need (much) larger sample sizes



Take care

- Frequency distributions with outliers (d) should not be analyzed with t-distribution



Parametric tests - Unequal variance

- Welch's *t*-test would work

- If sample sizes are equal and large (> 30) then even a ten-fold difference in variance (or a roughly 3-fold difference in SD) is *approximately* OK

What to do when the assumptions are not true

- If the sample sizes are large, and deviation from normality not too big, parametric tests may be OK anyway
- Transformations
- Non-parametric tests
- Randomization and resampling

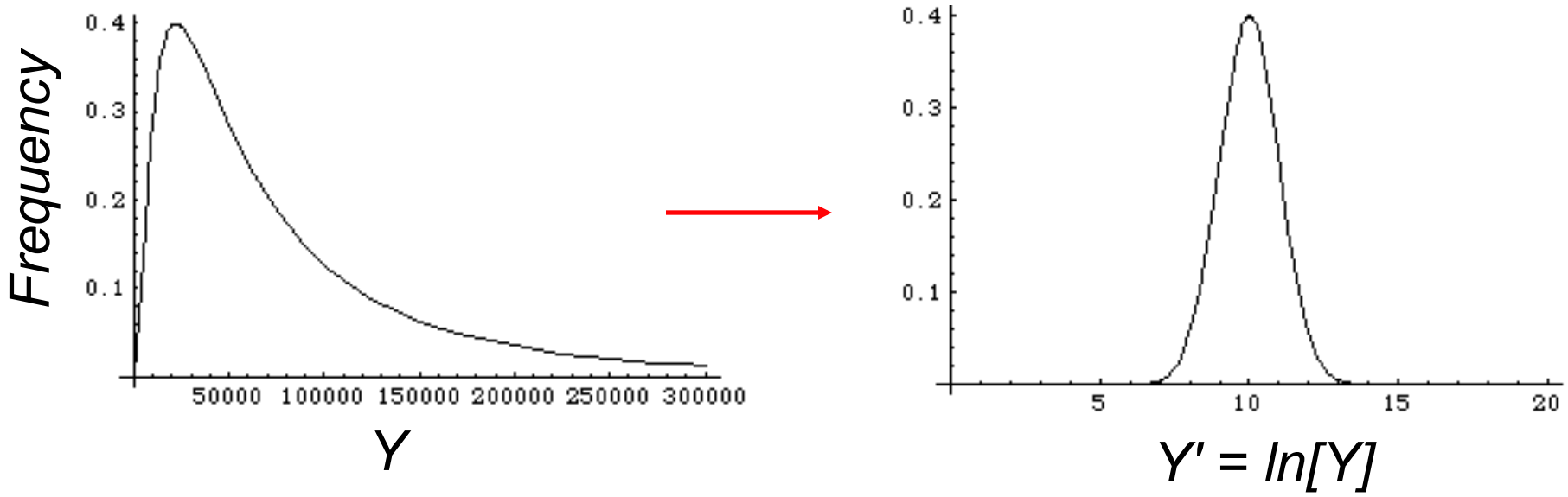
Data transformations

A data transformation changes each data point by some simple mathematical formula.

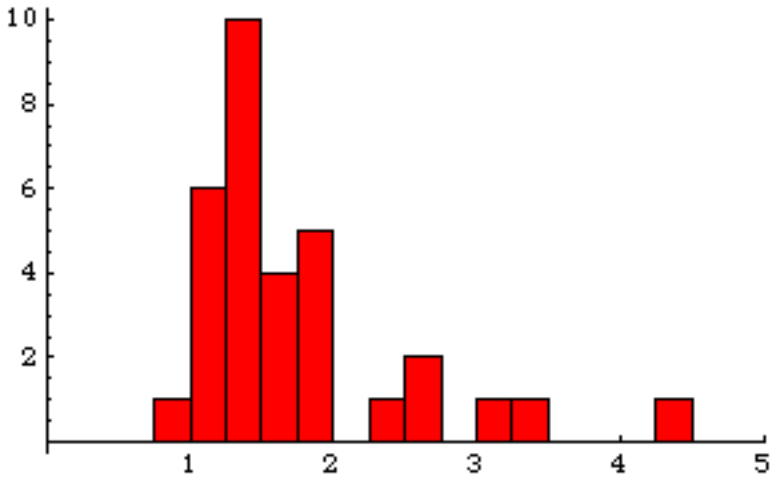
Log-transformation

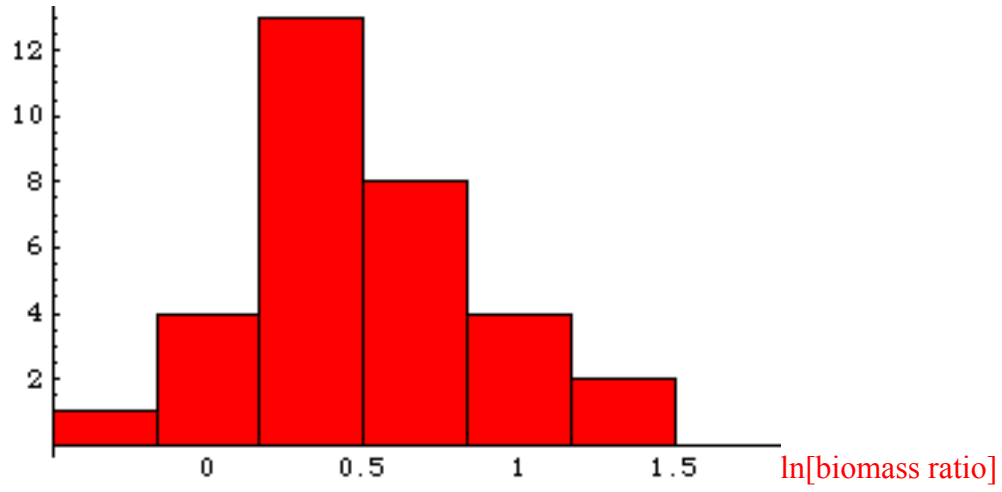
$$Y' = \ln[Y]$$

transformed



Biomass ratio	ln[Biomass Ratio]
1.34	0.30
1.96	0.67
2.49	0.91
1.27	0.24
1.19	0.18
1.15	0.14
1.29	0.26



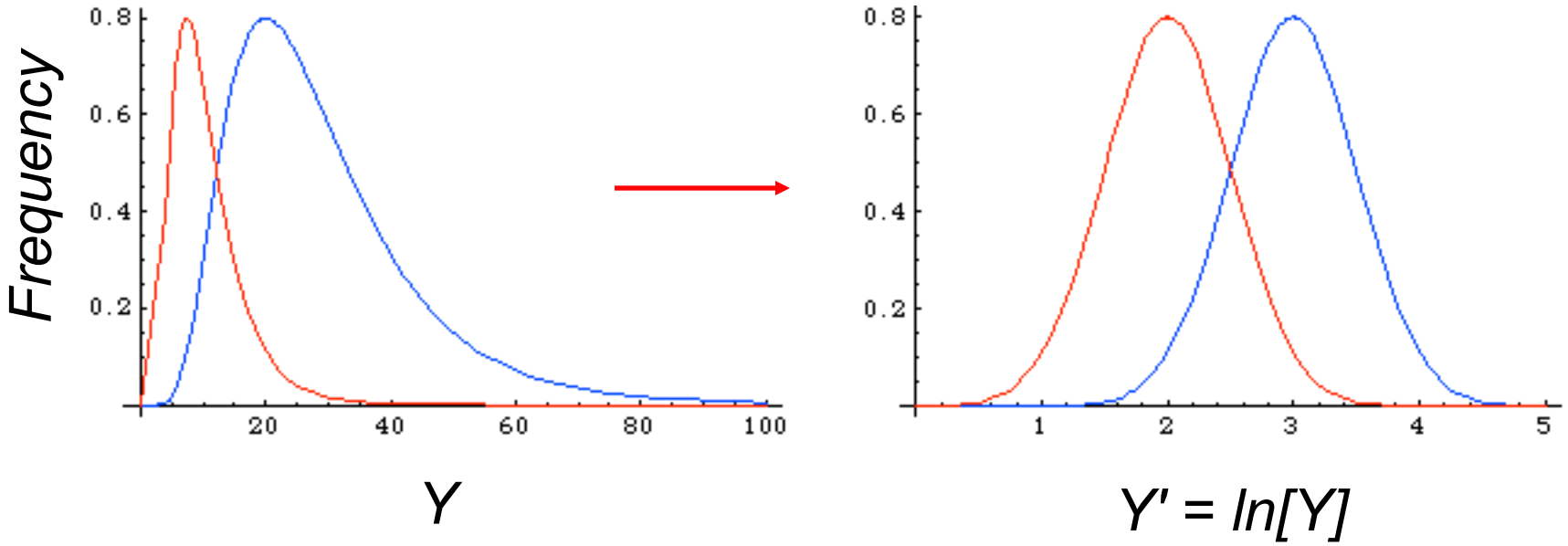


The transformed data are close to a normal distribution, allowing you to use a parametric test.

The log transformation is often useful when:

- *the variable is likely to be the result of multiplication of various components.*
- *the frequency distribution of the data is skewed to the right*
- *the variance increase as the mean gets larger (in comparisons across groups).*

Variance and mean increase together → try the log-transform



Other transformations

Arcsine	$p' = \arcsin\left[\sqrt{p}\right]$
Square-root	$Y' = \sqrt{Y + 1/2}$
Square	$Y' = Y^2$
Reciprocal	$Y' = \frac{1}{Y}$
Antilog	$Y' = e^Y$

Example: Confidence interval with log-transformed data

Data:	*5*	*12*	*1024*	*12398*
Log data:	*1.61*	*2.48*	*6.93*	*9.43*

$$\overline{Y}' = 5.11 \qquad s_{\ln[Y]} = 3.70$$

$$\overline{Y}' \pm t_{0.05(2),3}\frac{s_{\ln[Y]}}{\sqrt{n}} = 5.11 \pm 3.18\frac{3.70}{\sqrt{4}} = 5.11 \pm 5.88$$

$$-0.773 < \mu_{\ln[Y]} < 10.99$$

$$e^{-0.773} < e^{\mu_{\ln[Y]}} < e^{10.99}$$

$$0.46 < \mu_G < 59278$$

CI (for the geometric mean) on the original scale

Valid transformations...

- Require the same transformation be applied to each individual (and every group)
- Have one-to-one correspondence to original values (no ambiguity in transforming in either direction)
- Have a monotonic relationship with the original values (e.g., larger values stay larger)

Choosing transformations

- Must transform each individual in the same way
- You CAN try different transformations until you find one that makes the data fit the assumptions
- You CANNOT keep trying transformations until $P < 0.05$!!!

What to do when the assumptions are not true

- If the sample sizes are large, and deviation from normality not too big, parametric tests may be OK anyway
- Transformations
- Non-parametric tests
- Randomization and resampling

Non-parametric methods

- Assume less about the underlying distributions (so less power)
- Also called "distribution-free”

  (a bit of an exaggeration in some cases)
- "Parametric" methods assume a distribution or a parameter

Most non-parametric methods use **ranks**

- Rank each data point in all samples from lowest to highest

- Lowest data point gets rank 1, next lowest gets rank 2, ...

These tests, therefore, usually ask questions about medians, not means.

Sign test

- Non-parametric test
- Compares data from one sample to a constant
- Simple: for each data point, record whether individual is above (+) or below (-) the hypothesized constant
- Use a binomial test to compare result to *proportion* = 1/2
- Very little power

Sexual conflict and speciation

By Rickard Ignell, Swedish University of Agricultural Sciences - http://www.pheromone.ekol.lu.se/Images/mating.jpg from http://www.pheromone.ekol.lu.se/proj2camilla.html, CC BY-SA 1.0, https://commons.wikimedia.org/w/index.php?curid=6417662

Sexual conflict and speciation

- Is polygamy associated with higher or lower speciation rates?

Arnqvist et al. (2000) Sexual conflict promotes speciation in insects. PNAS 97:10460-10464.

Order	Family	Multiple mating group	Number of species	Single mating group	Number of species
Beetles	Anobiidae	Ernobius	53	Xestobium	10
	Dermestidae	*Dermestes*	73	*Trogoderma*	120
	Elateridae	*Agriotes*	228	*Selatosomus*	74
Flies	Muscidae	*Coenosia*	353	*Delia*	289
	Cecidomyiidae	*Rhopalomyia*	157	*Mayetiola*	30
	Chironomidae	*Chironomus*	300	*Pontomyia*	4
	Chironomidae	*Stictochironomus*	34	*Clunio*	18
	Drosophilidae and Culicidae	Drosophilidae	3,400	Culicidae	3,500
	Dryomyzidae and Calliphoridae	Dryomyzidae	20	Calliphoridae	1,000
	Tephritidae	*Anastrepha*	196	*Bactrocera*	486
	Sciaridae and Bibionidae	Sciaridae	1,750	Bibionidae	660
	Scatophagidae	*Scatophaga*	55	*Musca*	63
Mayflies	Siphlonuridae	*Siphlonurus*	37	*Caenis*	115
Homoptera	Psyllidae	*Cacopsylla*	100	*Aonidiella*	30
Butterflies and moths	Noctuidae and Psychidae	Noctuidae	21,000	Psychidae	600
	Tortricidae	*Choristoneura*	37	*Epiphyas*	40
	Nymphalidae	*Eueides (aliphera* clade*)*	7	*Eueides (vibilia* clade*)*	5
	Nymphalidae	*Heliconius (silvaniform* clade*)*	15	*Heliconius (sarasapho* clade*)*	7
	Nymphalidae	*Polygonia* /	18	*Nymphalis*	6

Etc....

Order	Family	Multiple mating group	Number of species	Single mating group	Number of species
Beetles	Anobiidae	Ernobius	53	Xestobium	10
	Dermestidae	*Dermestes*	73	*Trogoderma*	120

Paired data

(each group of multi mating matched with group of single mating)

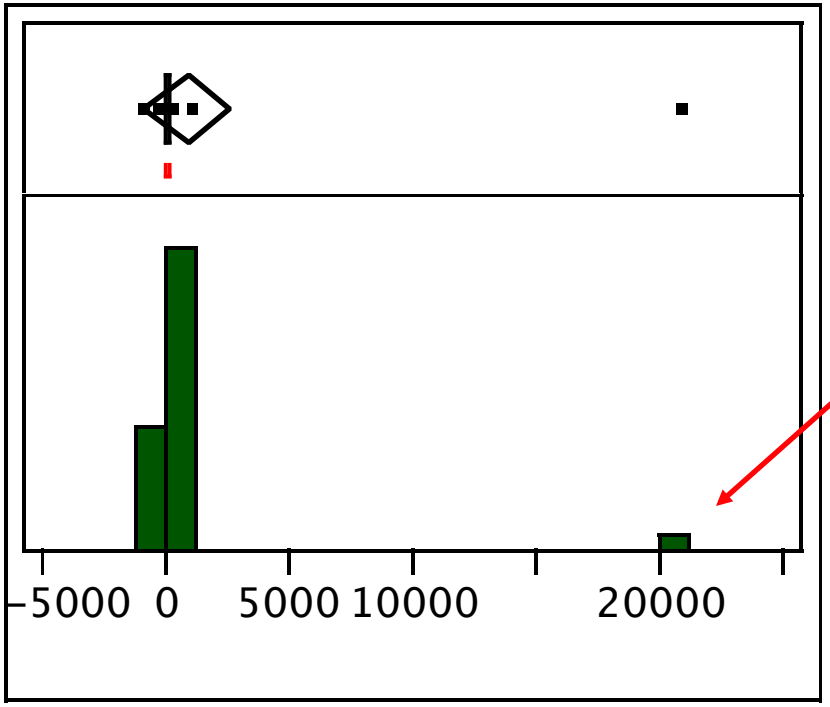
25 pairs

Get difference multiple – single mating

Make histogram

The differences are not normal

43	-47	154	64	127	296	16
-100	-980	-290	1090	-8	-78	70
20400	-3	2	8	12	227	1
61	1	79	78			



not normal
outlier

25 data points

Paired-t?

Can't rely on robustness

Hypotheses

H_0: The median difference in number of species between singly-mating and multiply-mating insect groups is 0.

H_A: The median difference in number of species between these groups is not 0.

7 out of 25 comparisons are negative

43	-47	154	64	127	296	16
-100	-980	-290	1090	-8	-78	70
20400	-3	2	8	12	227	1
61	1	79	78			

$$\Pr[X \le 7] = \sum_{i=0}^{7} \binom{25}{i} (0.5)^i (0.5)^{25-i} = 0.02164$$

P = 2 (0.02164) = 0.043

The sign test has very low power

What does that mean?

The sign test has very low power

So it is quite likely to *not* reject a *false* null hypothesis.

$$power = 1 - \beta$$
$$\beta = type\ II\ error$$

Non-parametric test to compare 2 groups

The Mann-Whitney U test compares the central tendencies of two groups using ranks.

Performing a Mann-Whitney U test

- First, rank all individuals from both groups together in order (for example, smallest to largest)

- Sum the ranks for all individuals in each group → R_1 and R_2

Calculating the test statistic, *U*

$$U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1$$

$$U_2 = n_1 n_2 - U_1$$

U_1 is the number of times an individual from pop. 1 has a lower rank than an individual from pop. 2, out of all pairwise comparisons.

Example: Garter snake resistance to newt toxin

Rough-skinned newt

Tetrodotoxin (TTX) in newt can harm snake

Comparing snake resistance to TTX (tetrodotoxin)

Locality	Resistance
Benton	0.29
Benton	0.77
Benton	0.96
Benton	0.64
Benton	0.70
Benton	0.99
Benton	0.34
Warrenton	0.17
Warrenton	0.28
Warrenton	0.20
Warrenton	0.20
Warrenton	0.37

This variable is known to be not normally distributed within populations.

Hypotheses

H_0: The TTX resistance for snakes from Benton is the same as for snakes from Warrenton.

H_A: The TTX resistance for snakes from Benton is different from snakes from Warrenton.

Calculating the ranks

Locality	Resistance	Rank
Benton	0.29	5
Benton	0.77	10
Benton	0.96	11
Benton	0.64	8
Benton	0.70	9
Benton	0.99	12
Benton	0.34	6
Warrenton	0.17	1
Warrenton	0.28	4
Warrenton	0.20	2.5
Warrenton	0.20	2.5
Warrenton	0.37	7

Rank sum for Warrenton: R=1+4+2.5+2.5+7=17

Calculating U_1 and U_2

$$U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1 = 5(7) + \frac{5(6)}{2} - 17 = 33$$

$$U_2 = n_1 n_2 - U_1 = 5(7) - 33 = 2$$

For a two-tailed test, we pick the larger of U_1 or U_2:

$$U = U_1 = 33$$

Compare U to the U table

(for alpha = 0.05)

n_1

n_2	3	4	5	6	7	8	9	10	11	12	13	14	15
3	–	–	15	17	20	22	25	27	30	32	35	37	40
4	–	16	19	22	25	28	32	35	38	41	44	47	50
5	15	19	23	27	30	34	38	42	46	49	53	57	61
6	17	22	27	31	36	40	44	49	53	58	62	67	71
7	20	25	30	36	41	46	51	56	61	66	71	76	81
8	22	28	34	40	46	51	57	63	69	74	80	86	91
9	25	32	38	44	51	57	64	70	76	82	89	95	101
10	27	35	42	49	56	63	70	77	84	91	97	104	111
11	30	38	46	53	61	69	76	84	91	99	106	114	121
12	32	41	49	58	66	74	82	91	99	107	115	123	131
13	35	44	53	62	71	80	89	97	106	115	124	132	141
14	37	47	57	67	76	86	95	104	114	123	132	141	151
15	40	50	61	71	81	91	101	111	121	131	141	151	161

Compare U to the U table

- Critical value for U for n_1 =5 and n_2=7 is 30
- 33 ≥ 30, so we can reject the null hypothesis
- Snakes from Benton have a different distribution of resistance to TTX than the Warrenton snakes.

How to deal with ties

- Determine the ranks that the values would have got if they were slightly different.
- Average these ranks, and assign that average to each tied individual
- Count all those individuals when deciding the rank of the next largest individual

Ties

Group	*Y*	*Rank*
2	12	1
2	14	2
1	17	3
1	19	4.5
2	19	4.5
1	24	6
2	27	7
1	28	8

Mann-Whitney: Large sample approximation

For n_1 and n_2 both greater than 10, use

$$Z = \frac{2U - n_1 n_2}{\sqrt{n_1 n_2 \left(n_1 + n_2 + 1 \right) / 3}}$$

Compare this Z to the standard normal distribution

Example:

U_1=245 U_2=80

n_1=13 n_2=25

$$Z = \frac{2U - n_1 n_2}{\sqrt{n_1 n_2 (n_1 + n_2 + 1)/3}}$$

$$= \frac{2(245) - 13(25)}{\sqrt{13(25)(13 + 25 + 1)/3}}$$

$$= 2.54$$

$Z_{0.05(2)}$=1.96, Z>1.96, so we can reject the null hypothesis.

Assumptions of Mann-Whitney U test

Both samples are random samples.

Both populations have the same shape of distribution.

For this last reason, this test is less useful than had originally been thought.

What to do when the assumptions are not true

- If the sample sizes are large, and deviation from normality not too big, parametric tests may be OK anyway
- Transformations
- Non-parametric tests
- Randomization and resampling

Permutation

A permutation test generates a null distribution for the association between two variables by repeatedly and randomly rearranging the values of one of the two variables in the data.

Make permutated data set

- Decide on test statistic (e.g. difference in mean between two groups)
- Reshuffle data

A	B
2.3	1.2
3.6	0.3
2.9	0.9
3.1	0.8

Make permutated data set

- Decide on test statistic (e.g. difference in mean between two groups)
- Reshuffle data

A	B
2.3	1.2
3.6	0.3
2.9	0.9
3.1	0.8

A	B
2.3	1.2
0.3	3.6
0.9	2.9
3.1	0.8

A	B
2.3	1.2
3.6	0.3
2.9	0.9
0.8	3.1

Calculate test stats and repeat

- Get the test stats for each permutation run
- Repeat at least 1000 times
- This set of test statistics of the randomized data sets is the null distribution
- What is the probability to get the test statistic (or more extreme) under this distribution?

Lets go through in example

- Two groups A and B
- Two samples (with very weird distributions)
- Test statistic is difference between mean (A – B)
- Let's derive a null distribution
- Compare the difference between the means of real data with the null distribution