NONPARAMETRIC TESTS I (DA_2022)

Andrea Giansanti

Dipartimento di Fisica, Sapienza Università di Roma

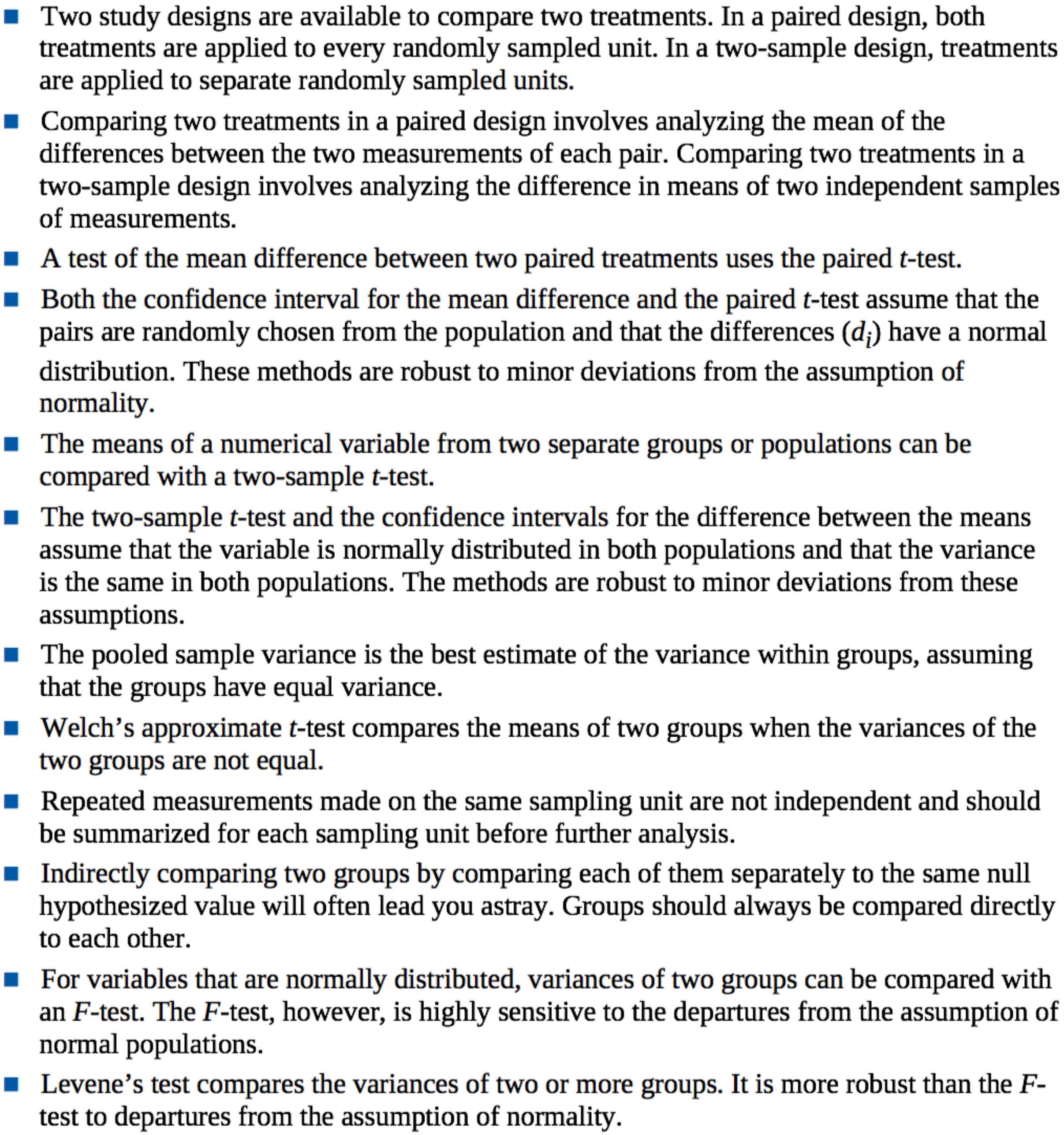
Andrea.Giansanti@roma1.infn.it

Lecture n. 20, Rome 8th of May 2022

DIPARTIMENTO DI FISICA

SAPIENZA
UNIVERSITÀ DI ROMA

SUMMARY OF TWO SAMPLE TESTS



- Two study designs are available to compare two treatments. In a paired design, both treatments are applied to every randomly sampled unit. In a two-sample design, treatments are applied to separate randomly sampled units.
- Comparing two treatments in a paired design involves analyzing the mean of the differences between the two measurements of each pair. Comparing two treatments in a two-sample design involves analyzing the difference in means of two independent samples of measurements.
- A test of the mean difference between two paired treatments uses the paired *t*-test.
- Both the confidence interval for the mean difference and the paired *t*-test assume that the pairs are randomly chosen from the population and that the differences (d_i) have a normal distribution. These methods are robust to minor deviations from the assumption of normality.
- The means of a numerical variable from two separate groups or populations can be compared with a two-sample *t*-test.
- The two-sample *t*-test and the confidence intervals for the difference between the means assume that the variable is normally distributed in both populations and that the variance is the same in both populations. The methods are robust to minor deviations from these assumptions.
- The pooled sample variance is the best estimate of the variance within groups, assuming that the groups have equal variance.
- Welch's approximate *t*-test compares the means of two groups when the variances of the two groups are not equal.
- Repeated measurements made on the same sampling unit are not independent and should be summarized for each sampling unit before further analysis.
- Indirectly comparing two groups by comparing each of them separately to the same null hypothesized value will often lead you astray. Groups should always be compared directly to each other.
- For variables that are normally distributed, variances of two groups can be compared with an *F*-test. The *F*-test, however, is highly sensitive to the departures from the assumption of normal populations.
- Levene's test compares the variances of two or more groups. It is more robust than the *F*-test to departures from the assumption of normality.

Outline DA_2022_L18

- The problem of non-normal distributed data
- Transformations: lognormal distributions
- Tests of normality
- Non parametric tests
- Sign test
- Binomial distribution

Study materials:

Rosner's chapter 9 and Whitlock's chap. 13 (very good)

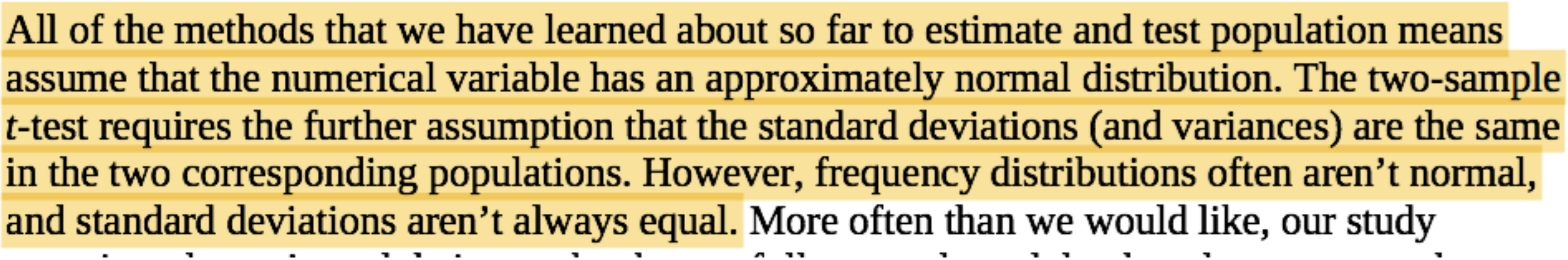
MOREOVER: As an exercise (to be recorded in the logbook) I suggest that you look at the very good scholarly lecture by professor Francesco Pauli of Trieste (in Italian) on the Neyman-Parson paradigm of testing hypotheses published on you tube https://www.youtube.com/watch?v=4jv7fKjn0Nc

TRANSLATE TO ENGLISH (committee)

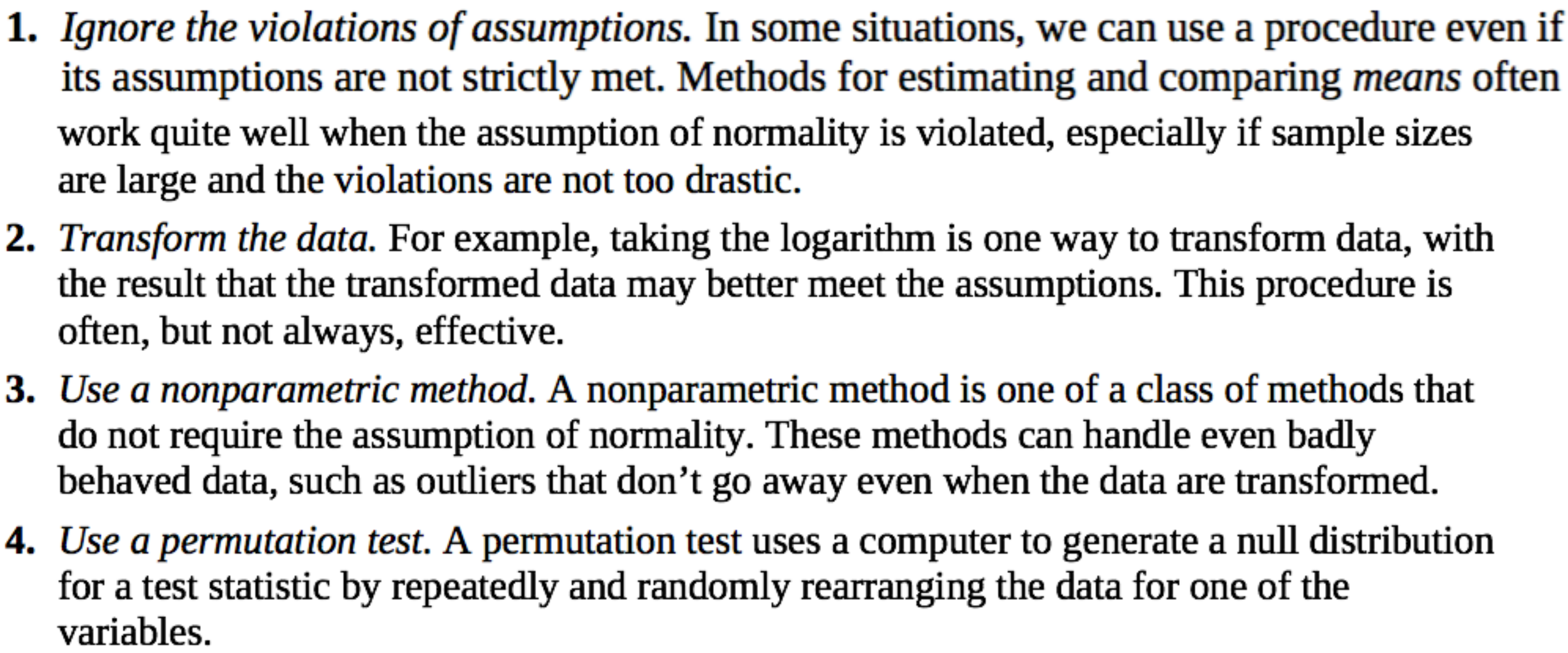
Consider the collection points of significance by Naomi Altman in Nature Methods

https://www.nature.com/collections/qghhqm/pointsofsignificance

THE PROBLEM OF DATA THAT VIOLATE THE REQUIREMENT OF NORMALITY



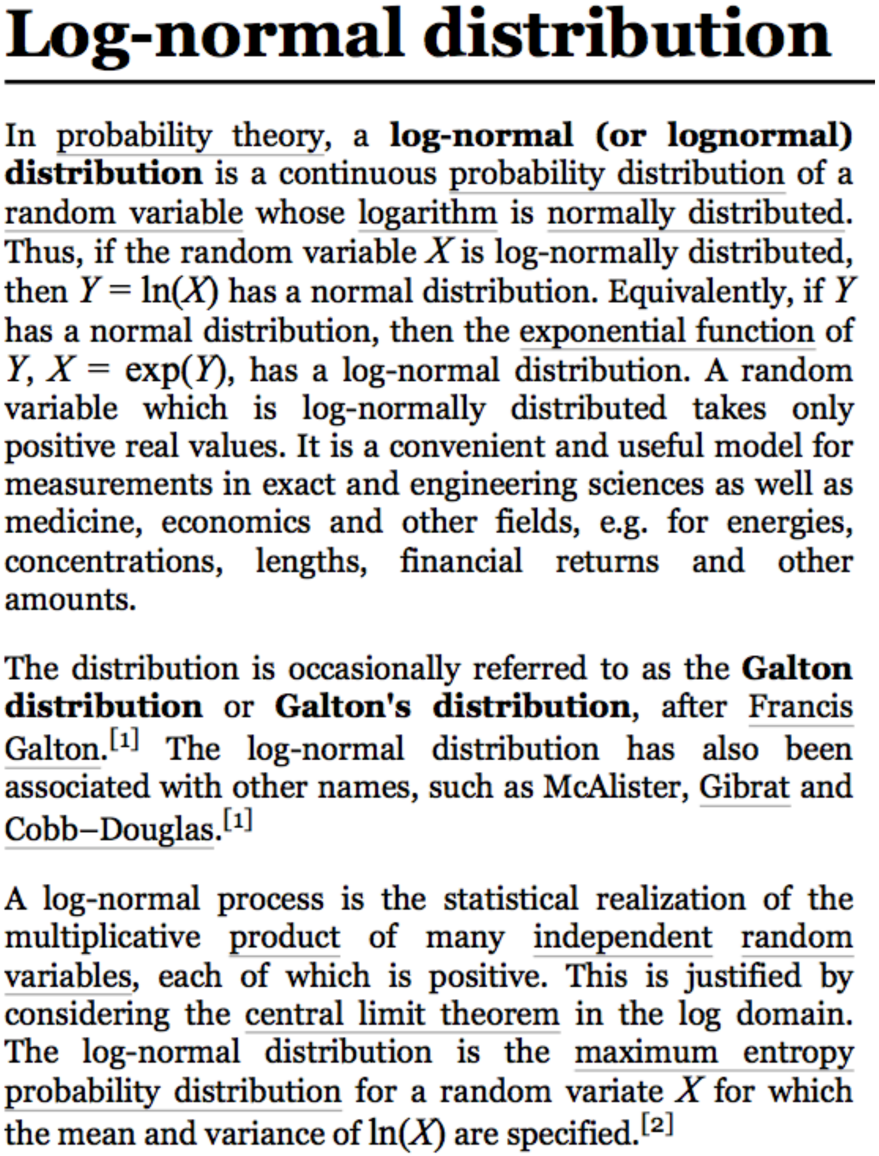
All of the methods that we have learned about so far to estimate and test population means assume that the numerical variable has an approximately normal distribution. The two-sample *t*-test requires the further assumption that the standard deviations (and variances) are the same in the two corresponding populations. However, frequency distributions often aren't normal, and standard deviations aren't always equal. More often than we would like, our study



1. *Ignore the violations of assumptions.* In some situations, we can use a procedure even if its assumptions are not strictly met. Methods for estimating and comparing *means* often work quite well when the assumption of normality is violated, especially if sample sizes are large and the violations are not too drastic.
2. *Transform the data.* For example, taking the logarithm is one way to transform data, with the result that the transformed data may better meet the assumptions. This procedure is often, but not always, effective.
3. *Use a nonparametric method.* A nonparametric method is one of a class of methods that do not require the assumption of normality. These methods can handle even badly behaved data, such as outliers that don't go away even when the data are transformed.
4. *Use a permutation test.* A permutation test uses a computer to generate a null distribution for a test statistic by repeatedly and randomly rearranging the data for one of the variables.

Michael C. Whitlock and Dolph Schluter - The Analysis of Biological Data-W. H. Freeman and Company (2015)

The fate of log-normal world: the distribution of multiplicative noise

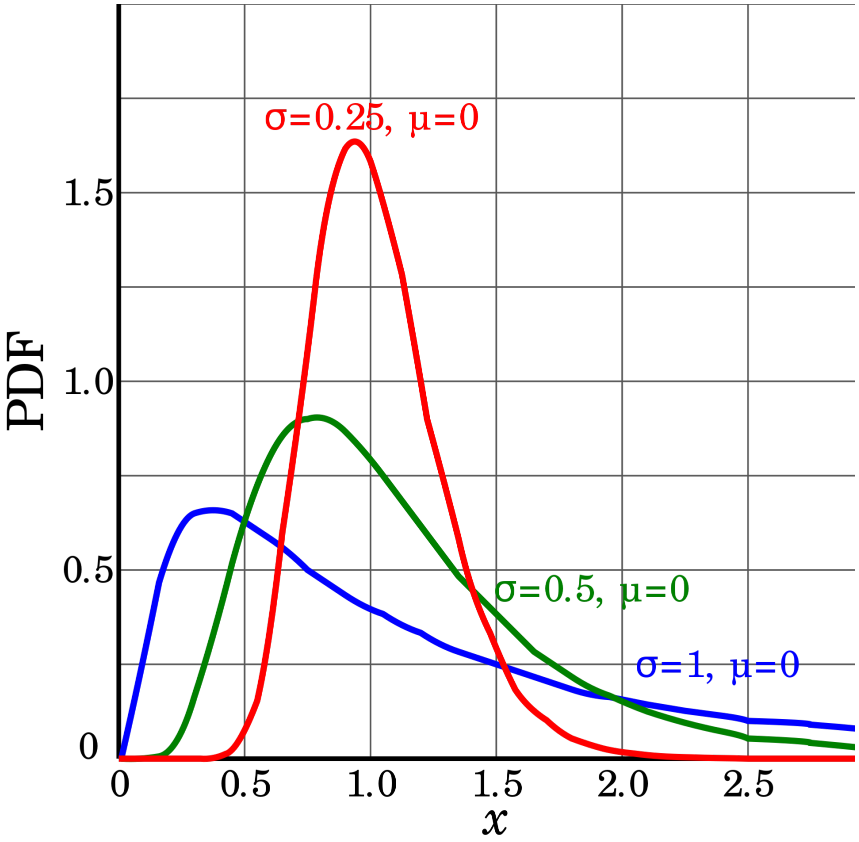


Log-normal distribution

In probability theory, a **log-normal (or lognormal) distribution** is a continuous probability distribution of a random variable whose logarithm is normally distributed. Thus, if the random variable X is log-normally distributed, then $Y = \ln(X)$ has a normal distribution. Equivalently, if Y has a normal distribution, then the exponential function of Y, $X = \exp(Y)$, has a log-normal distribution. A random variable which is log-normally distributed takes only positive real values. It is a convenient and useful model for measurements in exact and engineering sciences as well as medicine, economics and other fields, e.g. for energies, concentrations, lengths, financial returns and other amounts.

The distribution is occasionally referred to as the **Galton distribution** or **Galton's distribution**, after Francis Galton.[1] The log-normal distribution has also been associated with other names, such as McAlister, Gibrat and Cobb–Douglas.[1]

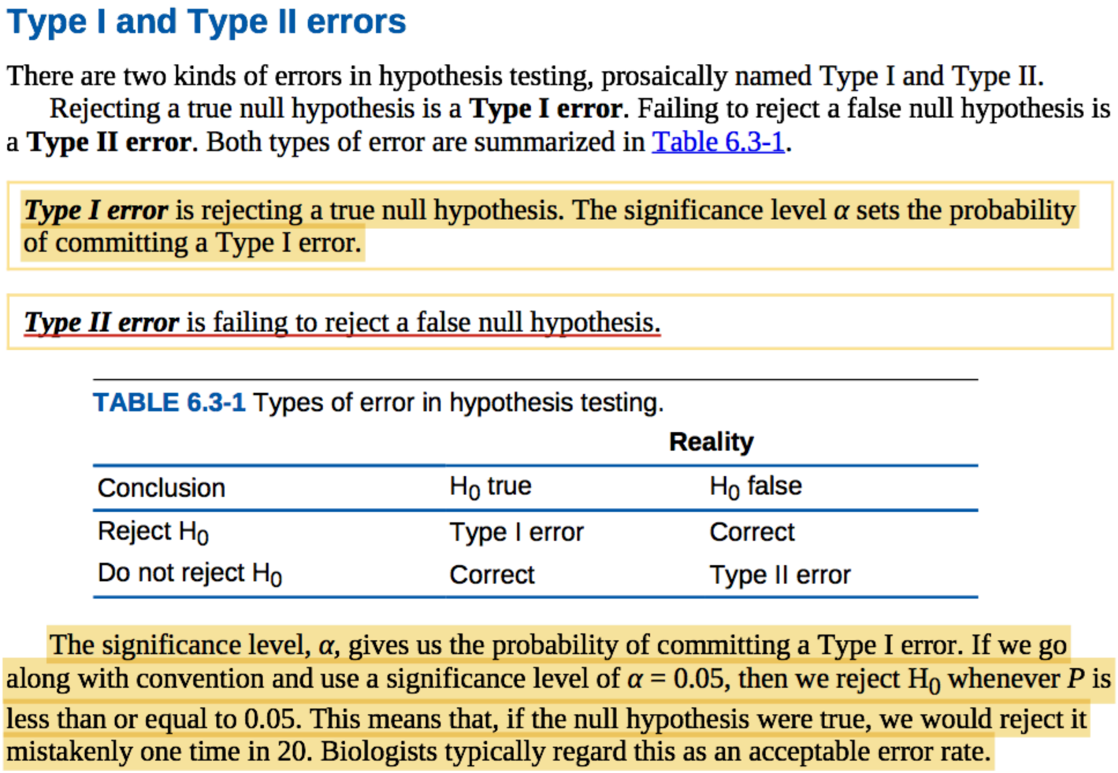
A log-normal process is the statistical realization of the multiplicative product of many independent random variables, each of which is positive. This is justified by considering the central limit theorem in the log domain. The log-normal distribution is the maximum entropy probability distribution for a random variate X for which the mean and variance of $\ln(X)$ are specified.[2]



DETECTING NON-NORMALITY

- Histograms (looking for skewness, asymmetry)
- Normal-quantile plots (normal probability plots)
- Shapiro-Wilk test (which has optimal power)

REM more Power of the test less risk of making type II errors



Type I and Type II errors

There are two kinds of errors in hypothesis testing, prosaically named Type I and Type II.

Rejecting a true null hypothesis is a **Type I error**. Failing to reject a false null hypothesis is a **Type II error**. Both types of error are summarized in Table 6.3-1.

Type I error is rejecting a true null hypothesis. The significance level α sets the probability of committing a Type I error.

Type II error is failing to reject a false null hypothesis.

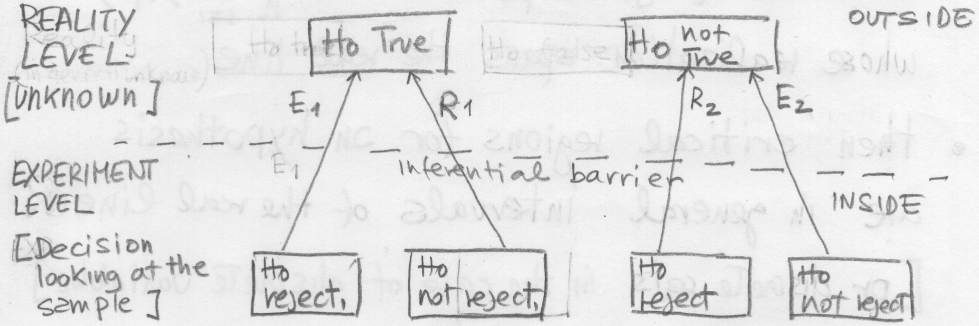
TABLE 6.3-1 Types of error in hypothesis testing.

	Reality	
Conclusion	H_0 true	H_0 false
Reject H_0	Type I error	Correct
Do not reject H_0	Correct	Type II error

The significance level, α, gives us the probability of committing a Type I error. If we go along with convention and use a significance level of $\alpha = 0.05$, then we reject H_0 whenever P is less than or equal to 0.05. This means that, if the null hypothesis were true, we would reject it mistakenly one time in 20. Biologists typically regard this as an acceptable error rate.

IMPORTANT REMARKS

- H_0 is either true or false (tertium non datum)



We are in a probabilistic setting; we do not know if we are taking the right or the wrong choice. E_1 is a type I ERROR (H_0 rejected when it is true); E_2 is a type II ERROR (H_0 not rejected when it is not true). Then let us define:

$$\alpha = P(E_1) = P(\text{reject } H_0 \mid H_0 \text{ true}) = P(\underline{X} \in C_0 \mid H_0)$$

$$\beta = P(E_2) = P(\text{not reject } H_0 \mid H_0 \text{ false}) = P(\underline{X} \notin C_0 \mid H_1)$$

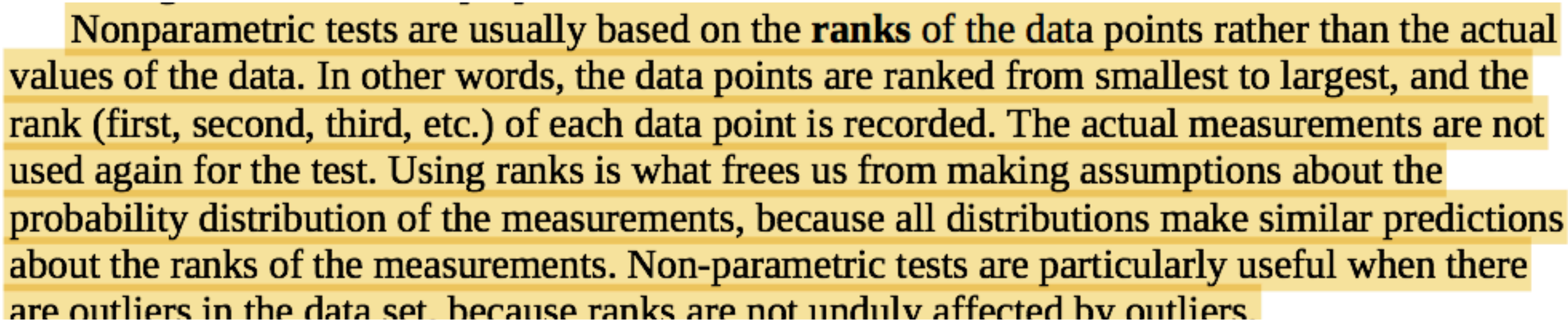
$$1-\alpha = P(R_1) = P(\text{not reject } H_0 \mid H_0 \text{ true}) = P(\underline{X} \notin C_0 \mid H_0)$$

$$\gamma = 1-\beta = P(R_2) = P(\text{reject } H_0 \mid H_0 \text{ false}) = P(\underline{X} \in C_0 \mid H_1)$$

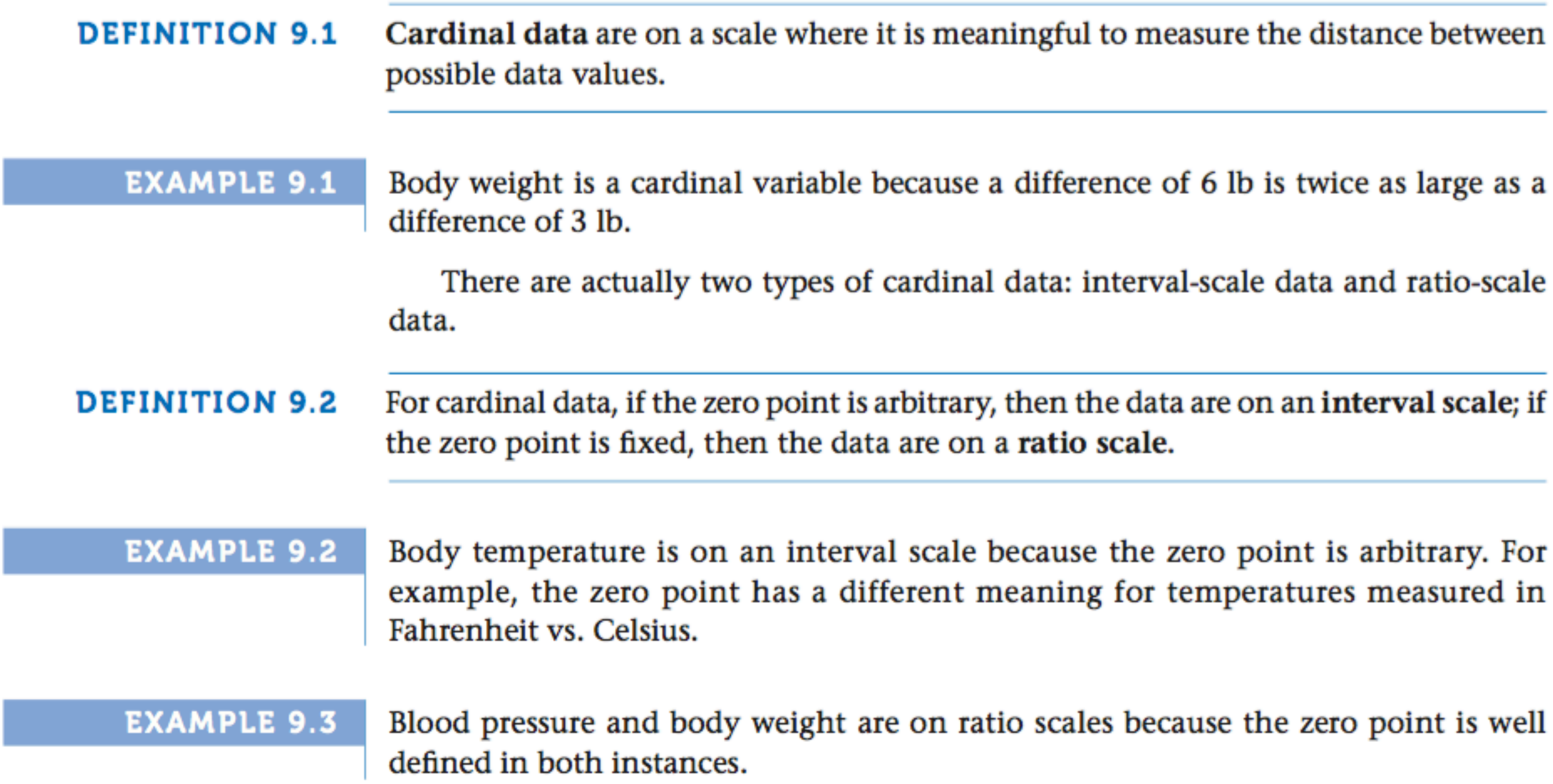
α is also called SIGNIFICANCE of a TEST

γ is also called POWER of a TEST

What is a Rank? How to rank data?



Nonparametric tests are usually based on the **ranks** of the data points rather than the actual values of the data. In other words, the data points are ranked from smallest to largest, and the rank (first, second, third, etc.) of each data point is recorded. The actual measurements are not used again for the test. Using ranks is what frees us from making assumptions about the probability distribution of the measurements, because all distributions make similar predictions about the ranks of the measurements. Non-parametric tests are particularly useful when there are outliers in the data set, because ranks are not unduly affected by outliers.



DEFINITION 9.1 **Cardinal data** are on a scale where it is meaningful to measure the distance between possible data values.

EXAMPLE 9.1 Body weight is a cardinal variable because a difference of 6 lb is twice as large as a difference of 3 lb.

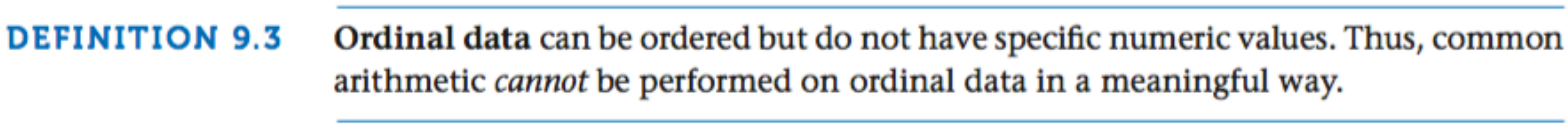
There are actually two types of cardinal data: interval-scale data and ratio-scale data.

DEFINITION 9.2 For cardinal data, if the zero point is arbitrary, then the data are on an **interval scale**; if the zero point is fixed, then the data are on a **ratio scale**.

EXAMPLE 9.2 Body temperature is on an interval scale because the zero point is arbitrary. For example, the zero point has a different meaning for temperatures measured in Fahrenheit vs. Celsius.

EXAMPLE 9.3 Blood pressure and body weight are on ratio scales because the zero point is well defined in both instances.

ORDINAL DATA, PRE_measurements

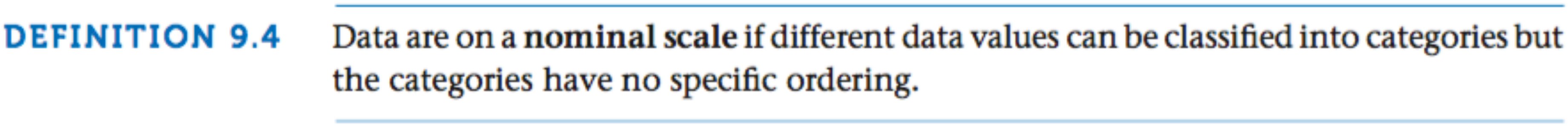


DEFINITION 9.3 **Ordinal data** can be ordered but do not have specific numeric values. Thus, common arithmetic *cannot* be performed on ordinal data in a meaningful way.

AT THIS POINT DISCUSS ON THE IPAD THE OPERATIONAL DEFINITION OF WEIGHT:

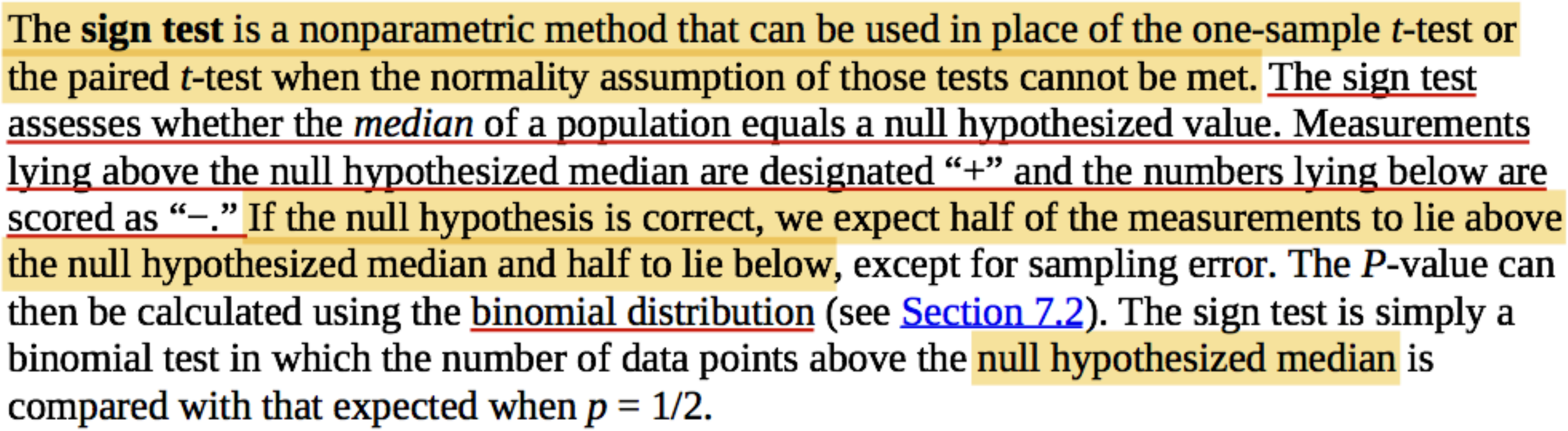
USING A PRE-BALANCE and then ADDING A SCALE (measurement unit)

REM DATA CAN BE CARDINAL, ORDINAL, CATEGORICAL (NOMINAL)

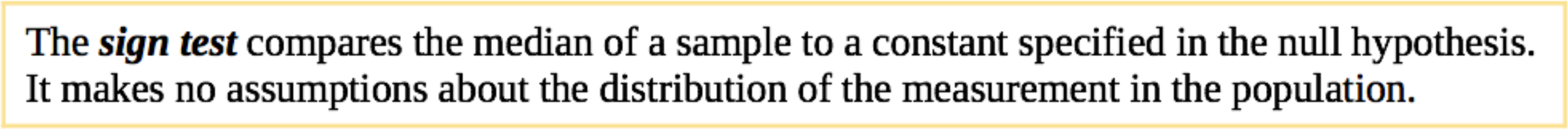


DEFINITION 9.4 Data are on a **nominal scale** if different data values can be classified into categories but the categories have no specific ordering.

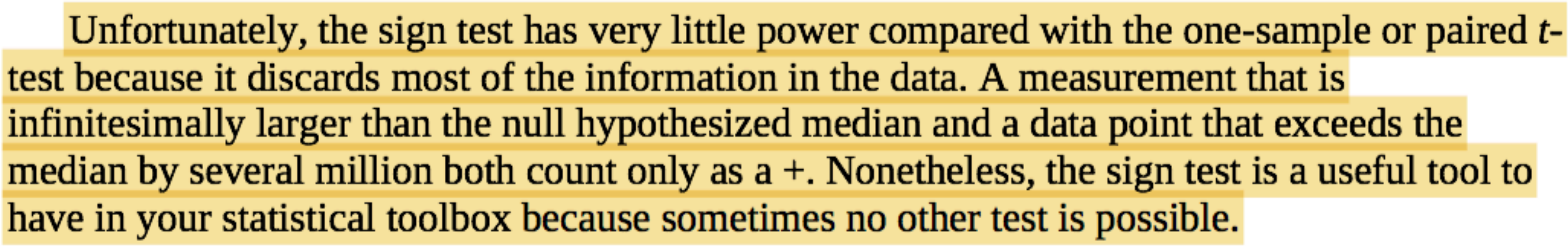
Sign test



The **sign test** is a nonparametric method that can be used in place of the one-sample *t*-test or the paired *t*-test when the normality assumption of those tests cannot be met. The sign test assesses whether the *median* of a population equals a null hypothesized value. Measurements lying above the null hypothesized median are designated "+" and the numbers lying below are scored as "−." If the null hypothesis is correct, we expect half of the measurements to lie above the null hypothesized median and half to lie below, except for sampling error. The *P*-value can then be calculated using the binomial distribution (see Section 7.2). The sign test is simply a binomial test in which the number of data points above the null hypothesized median is compared with that expected when $p = 1/2$.



The ***sign test*** compares the median of a sample to a constant specified in the null hypothesis. It makes no assumptions about the distribution of the measurement in the population.



Unfortunately, the sign test has very little power compared with the one-sample or paired *t*-test because it discards most of the information in the data. A measurement that is infinitesimally larger than the null hypothesized median and a data point that exceeds the median by several million both count only as a +. Nonetheless, the sign test is a useful tool to have in your statistical toolbox because sometimes no other test is possible.

EXAMPLE 13.4 Sexual conflict and the origin of new species

Horia Bogdan/Shutterstock.com

The process by which a single species splits into two species is still not well understood. One proposal involves "sexual conflict" – a genetic arms race between males and females that arises from their different reproductive roles.[8] Sexual conflict can cause rapid genetic divergence between isolated populations of the same species, leading to the formation of new species. Sexual conflict is more pronounced in species in which females mate more than once, leading to the prediction that they should form new species at a more rapid rate. To investigate this, Arnqvist et al. (2000) identified 25 insect taxa (groups) in which females mate multiple times, and they paired each of these groups to a closely related insect group in which females only mate once. Which type of insect tends to have more species? Table 13.4-1 lists the numbers of insect species in each of the groups.

TABLE 13.4-1 The number of species in 25 pairs of insect groups. Each pair matches a group of insect species in which females mate only once with a related group of insect species in which females mate multiple times.

	Number of species			
Taxon pair	Multiple-mating group	Single-mating group	Difference	Above (+) or below (−) zero
A	53	10	43	+
B	73	120	−47	−
C	228	74	154	+
D	353	289	64	+
E	157	30	127	+
F	300	4	296	+
G	34	18	16	+
H	3400	3500	−100	−
I	20	1000	−980	−
J	196	486	−290	−
K	1750	660	1090	+
L	55	63	−8	−
M	37	115	−78	−
N	100	30	70	+
O	21,000	600	20,400	+
P	37	40	−3	−
Q	7	5	2	+
R	15	7	8	+
S	18	6	12	+
T	240	13	227	+
U	15	14	1	+
V	77	16	61	+
W	15	14	1	+
X	85	6	79	+
Y	86	8	78	+

The data are paired. Thus, for each group of insects whose females mate once, there is a corresponding, closely related group of insect species in which females mate more than once. For this reason, the analysis must focus on the paired differences. The differences listed in Table 13.4-1 were calculated by subtracting the number of species in the single-mating group from that of the corresponding multiple-mating group.

The data are paired. Thus, for each group of insects whose females mate once, there is a corresponding, closely related group of insect species in which females mate more than once. For this reason, the analysis must focus on the paired differences. The differences listed in Table 13.4-1 were calculated by subtracting the number of species in the single-mating group from that of the corresponding multiple-mating group.

First, examine the histogram of the differences in Figure 13.4-1. These data have one outlier at 20,400, and we hardly need a Shapiro-Wilk test (Section 13.1) to tell us that the measurements are not normally distributed. At the same time, there are only 25 data points, which is too small a sample size to rely on the robustness of the paired *t*-test. There is no obvious transformation that would make these data normal, so we should pursue a nonparametric test instead. We will use the sign test to evaluate whether the median of the difference equals zero.

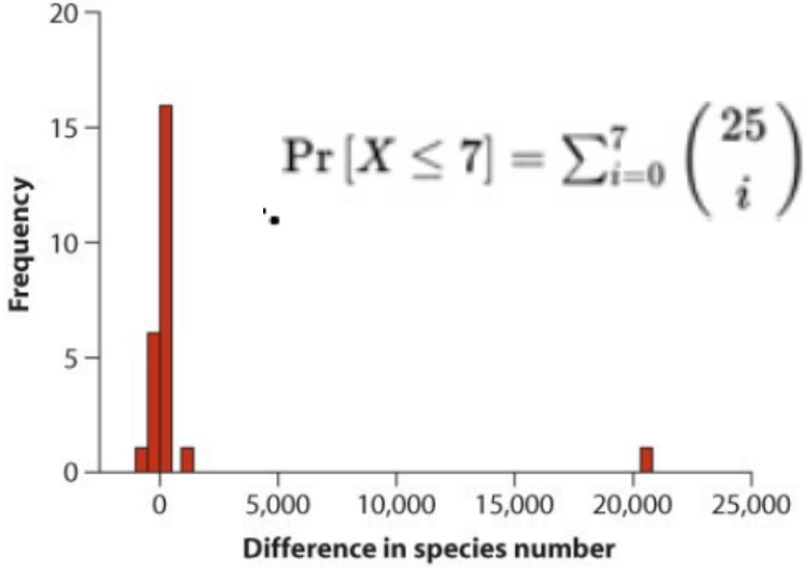


Figure 13.4-1
Whitlock et al., *The Analysis of Biological Data*, 2e, © 2015 W. H. Freeman and Company

FIGURE 13.4-1 The distribution of differences in species number between single-mating and multiple-mating insect groups. There is an extreme outlier at 20,400.

Our hypotheses are as follows.

H_0: The median difference in number of species between insect groups is zero.

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

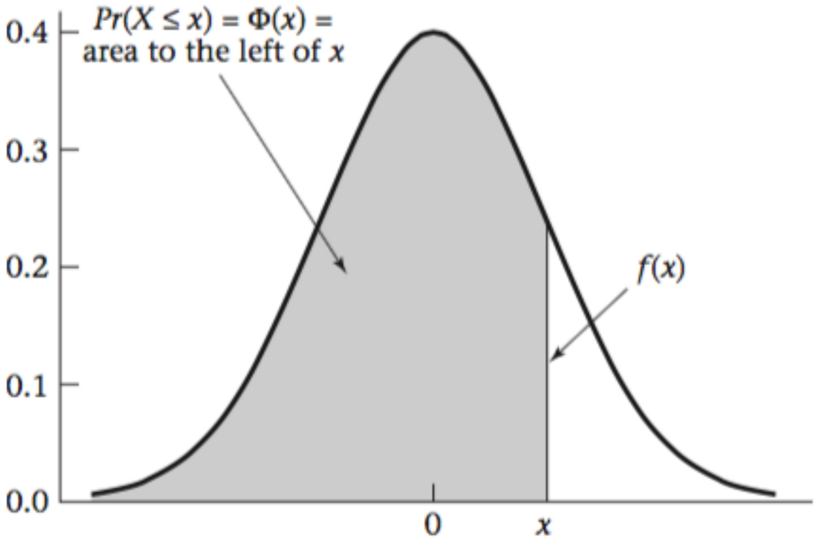
From this point on, the sign test is the same as the binomial test. If the null hypothesis is correct, then we expect half the measurements to fall above zero (+) and half to fall below zero (−). In fact, 18 out of the 25 measurements fall above zero and only seven fall below (see the last column in Table 13.4-1).

We can use the binomial distribution to calculate the *P*-value for the test. What is the probability of getting seven or fewer "−" observations out of 25 when the probability of a "−" observation is 0.5 under the null hypothesis? The answer is

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

The sign test is actually a special case of the one-sample binomial test in Section 7.9, where the hypothesis H_0: $p = 1/2$ vs. H_1: $p \neq 1/2$ was tested. In Equation 9.1 and Equation 9.2 a large-sample test is being used, and we are assuming the normal approximation to the binomial distribution is valid. Under H_0, $p = 1/2$ and $E(C) = np = n/2$, $Var(C) = npq = n/4$, and $C \dot{\sim} N(n/2, n/4)$. Furthermore, the .5 term in computing the critical region and p-value serves as a continuity correction and better approximates the binomial distribution by the normal distribution.

The cdf [$\Phi(x)$] for a standard normal distribution



The Sign Test

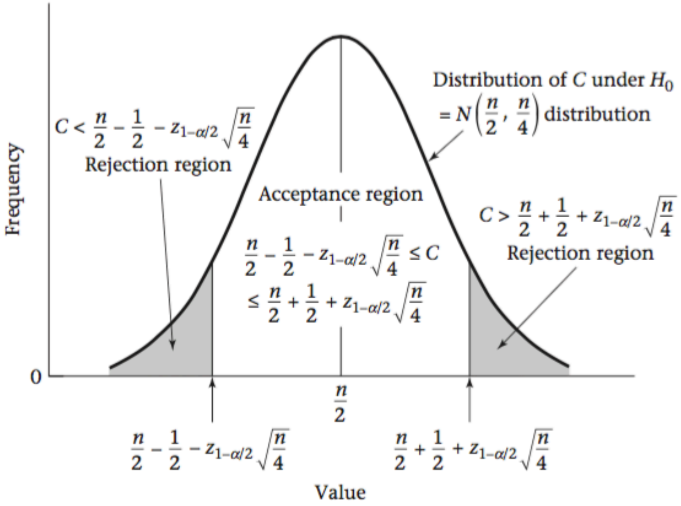
To test the hypothesis H_0: $\Delta = 0$ vs. H_1: $\Delta \neq 0$ with type I error = α, where the number of nonzero d_i's = $n \geq 20$ and C = the number of d_i's where $d_i > 0$, if

$$C > c_2 = \frac{n}{2} + \frac{1}{2} + z_{1-\alpha/2}\sqrt{n/4} \quad \text{or} \quad C < c_1 = \frac{n}{2} - \frac{1}{2} - z_{1-\alpha/2}\sqrt{n/4}$$

then H_0 is rejected. Otherwise, H_0 is accepted.

The acceptance and rejection regions for this test are shown in Figure 9.1.

Acceptance and rejection regions for the sign test



Similarly, the p-value for the procedure is computed using the following formula.

Computation of the p-Value for the Sign Test (Normal-Theory Method)

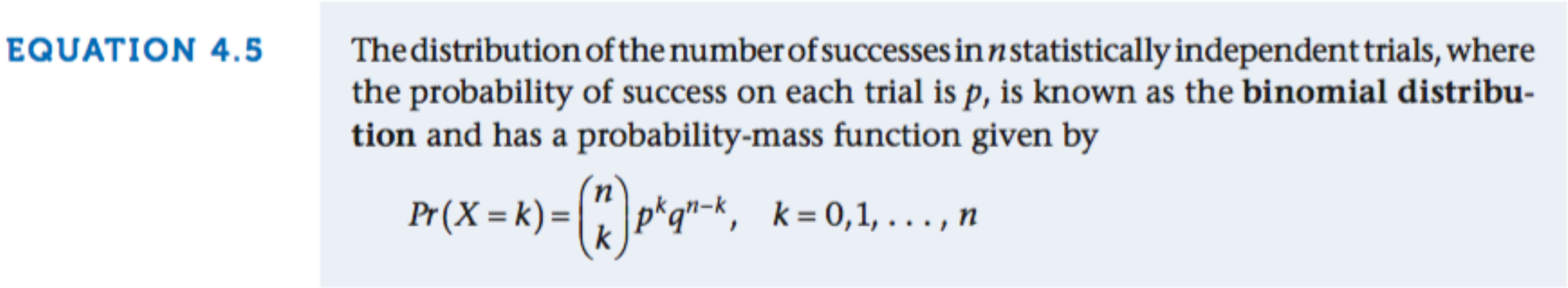
$$p = 2 \times \left[1 - \Phi\left(\frac{C - \frac{n}{2} - .5}{\sqrt{n/4}}\right)\right] \quad \text{if} \quad C > \frac{n}{2}$$

$$p = 2 \times \Phi\left(\frac{C - \frac{n}{2} + .5}{\sqrt{n/4}}\right) \quad \text{if} \quad C < \frac{n}{2}$$

$$p = 1.0 \quad \text{if} \quad C = \frac{n}{2}$$

This computation is illustrated in Figure 9.2.

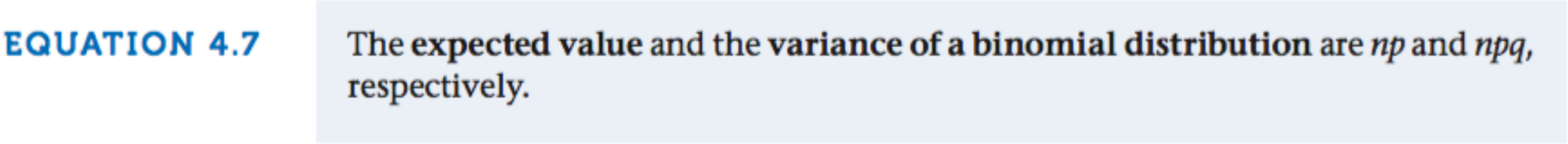
THE BINOMIAL DISCRETE DISTRIBUTION



EQUATION 4.5

The distribution of the number of successes in n statistically independent trials, where the probability of success on each trial is p, is known as the **binomial distribution** and has a probability-mass function given by

$$Pr(X=k)=\binom{n}{k}p^k q^{n-k}, \quad k=0,1,\ldots,n$$



EQUATION 4.7

The **expected value** and the **variance of a binomial distribution** are np and npq, respectively.

FIGURE 4.4 Plot of *pq* versus *p*

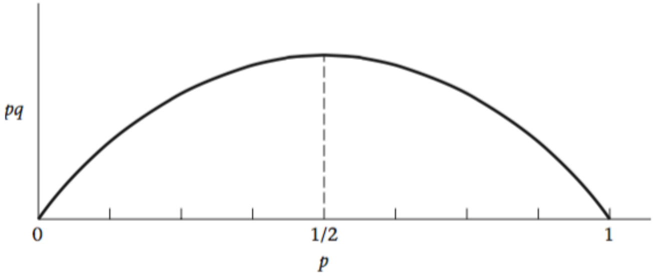
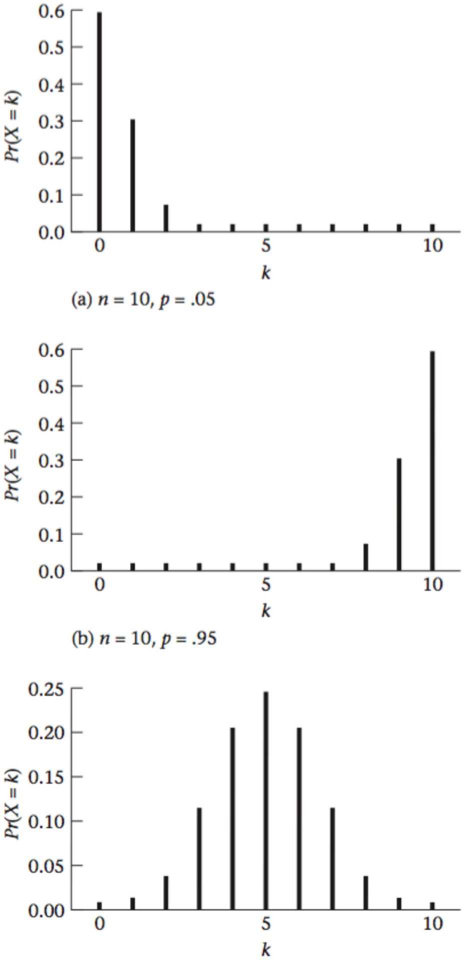


FIGURE 4.5 The binomial distribution for various values of *p* when *n* = 10



(a) $n = 10$, $p = .05$

(b) $n = 10$, $p = .95$

(c) $n = 10$, $p = .50$