Assessing Normality in a Sample

BIOS 6611

CU Anschutz

Week 3

1. **Assessing Normality**

2. **Graphical Approaches**

3. **Formal Statistical Tests**

Assessing Normality

Why Assess Normality?

Oftentimes we will see a statistical test has an assumption that the data are normally distributed. If that assumption is violated, the test may not perform as expected (e.g., worse power, biased results, etc.).

However, different tests have varying robustness to departures from normality (i.e., we do not always have to use alternative methods depending on how non-normal the data is!).

We can assess normality holistically with both graphical and statistical approaches.

Things to Keep in Mind

In general, when evaluating if an empirical (sample) distribution is normal:

1. Does the mean $=$ median $=$ mode? This is true for any symmetrical distribution, so it is not enough to determine normality alone, but can alert you to data that is not normal.
2. Is the data skewed? We know that a normal distribution should have a skewness of 0. Values $>$0 indicate it is skewed to the right ($+$), whereas $<$0 indicates skewed to the left (-).
3. What do the tails look like (kurtosis)? A true normal distribution will have a kurtosis of 3 (i.e., an *excess* kurtosis of 0).

Our Simulated Data

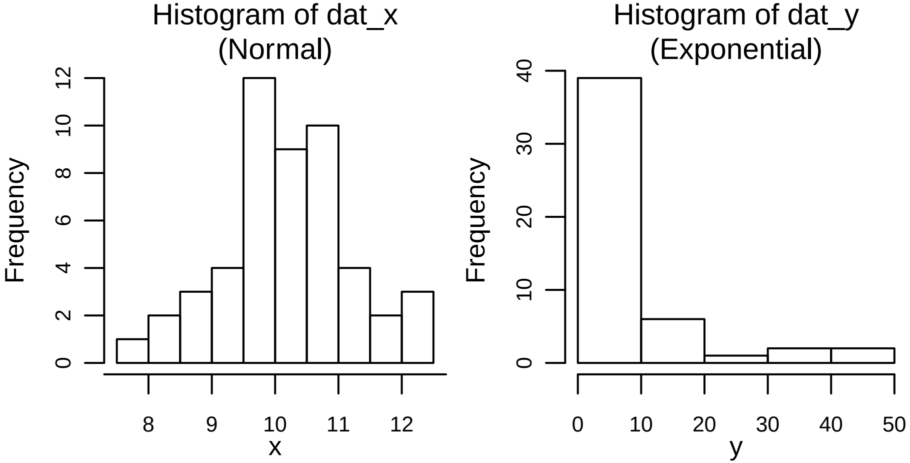
For this lecture, we will simulate data from a normal and exponential distribution with mean 10 so that $X \sim N(\mu = 10, \sigma^2 = 1)$ and $Y \sim Exp(\lambda = 0.1)$:

```r
set.seed(515)

dat_x <- rnorm(50, mean=10, sd=1)

dat_y <- rexp(50, rate=0.1) # the mean is 1/rate
```

Graphical Approaches

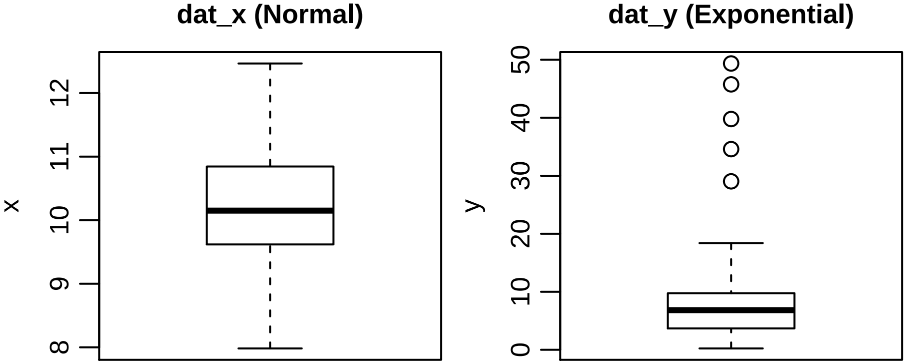
Histograms

A histogram groups continuous data into ranges and represents the frequency of a range by the height of the bar. Do we see symmetry? Do we see heavy or light tails?



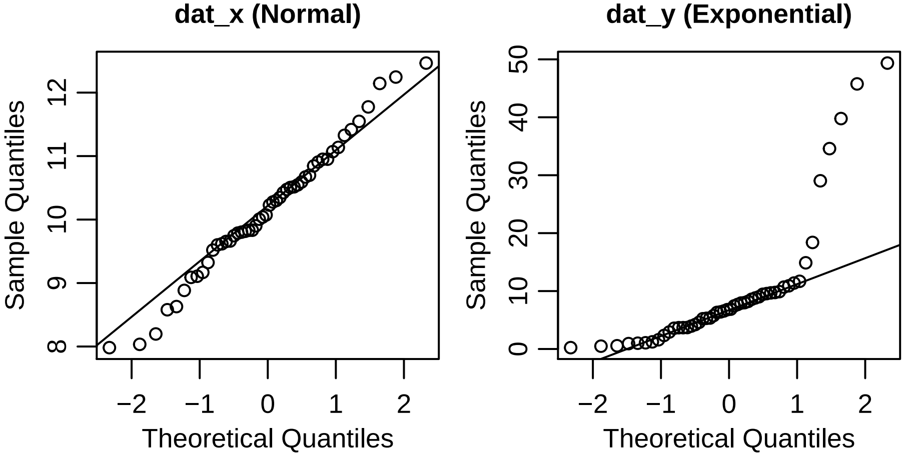
Box Plots

A box plot is useful as it usually shows the 1 quartile, median, and 3rd quartile in the "box", and then whiskers that extend to either the min/max or, if outliers are present, stop at a predefined calculation and indicate extreme values with points. In R we can use `boxplot()`.



Normal Probability Plots (Q-Q Plots)

Another plot that we can look at is the normal probability plot. It shows the observed percentiles from the data versus expected (theoretical) percentiles of a normal distribution. Deviations from a straight line may suggest departures from normality. (R functions: `qqnorm` and `qqline`.)



Formal Statistical Tests

Formal Statistical Tests

There are a number of statistical tests that look for deviations from normality: Shapiro-Wilk, Anderson-Darling, Cramer-von Mises, Lilliefors/Kolmogorov-Smirnov, etc.

Attached to most of these tests is a probability that you'd see a value of that statistic or one bigger (or one smaller for the Shapiro Wilk test) if the data are really, really normal, i.e. a p-value of this test.

It's best not to weight the results of these tests too heavily - small sample sizes can lead to a decision of normality even when the data are not normal and large sample sizes can lead to a decision of non-normality even when the data are very close to normal.

R Code

```r
library(nortest)

TestNormality <- function(x) {
    st <- shapiro.test(x)
    ks <- lillie.test(x)
    cvm <- cvm.test(x)
    ad <- ad.test(x)
    r1 <- c("W", round(st$statistic, 3), round(st$p.value, 3))
    r2 <- c("D", round(ks$statistic, 3), round(ks$p.value, 3))
    r3 <- c("W", round(cvm$statistic, 3), round(cvm$p.value, 3))
    r4 <- c("A", round(ad$statistic, 3), round(ad$p.value, 3))
    out <- data.frame(rbind(r1, r2, r3, r4))
    names(out) <- c("statistic", "value", "p.value")
    row.names(out) <- c("Shapiro-Wilk",
    "Lillies (Kolmogorov-Smirnov)", "Cramer-von Mises","Anderson-Darling")
    results <- out
    return(results)
}

TestNormality(dat_x); TestNormality(dat_y)
```

Results

Notice the difference in statistic values between our normal and exponential data sets, and how Shapiro-Wilk is significant for smaller W statistics.

Normality Test (Statistic)	**Function in R (Package)**	**dat_x Statistic; p-value**	**dat_y Statistic; p-value**
Shapiro-Wilk (W)	shapiro.test (stats package)	W=0.987; p=0.846	W=0.678; p<0.001
Anderson-Darling (D)	ad.test (nortest package)	D=0.073; p=0.730	D=0.289; p<0.001
Cramer-von Mises (W)	cvm.test (nortest package)	W=0.028; p=0.871	W=1.025; p<0.001
Lilliefors/ Kolmogorov-Smirnov (A)	lillie.test (nortest package)	A=0.192; p=0.892	A=5.653; p<0.001

In Conclusion. . .

THE
NORMAL
LAW OF ERROR
STANDS OUT IN THE
EXPERIENCE OF MANKIND
AS ONE OF THE BROADEST
GENERALIZATIONS OF NATURAL
PHILOSOPHY - IT SERVES AS THE
GUIDING INSTRUMENT IN RESEARCHES
IN THE PHYSICAL AND SOCIAL SCIENCES AND
IN MEDICINE AGRICULTURE AND ENGINEERING.
IT IS AN INDISPENSABLE TOOL FOR THE ANALYSIS AND THE
INTERPRETATION OF THE DATA OBTAINED BY OBSERVATION AND EXPERIMENT
W.J. YOUDEN