Chapter 4

Checking Assumptions

Learning objectives

After completing this topic, you should be able to:

assess the assumptions visually and via formal tests.

Achieving these goals contributes to mastery in these course learning outcomes:

10. Model assumptions.

4.1 Introduction

Almost all statistical methods make assumptions about the data collection process and the shape of the population distribution. If you reject the null hypothesis in a test, then a reasonable conclusion is that the null hypothesis is false, provided all the distributional assumptions made by the test are satisfied. If the assumptions are not satisfied then that alone might be the cause of rejecting H_0. Additionally, if you fail to reject H_0, that could be caused solely by failure to satisfy assumptions also. Hence, you should always check assumptions to the best of your abilities.

Two assumptions that underly the tests and CI procedures that I have discussed are that the data are a random sample, and that the population frequency curve is normal. For the pooled variance two-sample test the population variances are also required to be equal.

The random sample assumption can often be assessed from an understanding of the data collection process. Unfortunately, there are few general tests for checking this assumption. I have described exploratory (mostly visual) methods to assess the normality and equal variance assumption. I will now discuss formal methods to assess these assumptions.

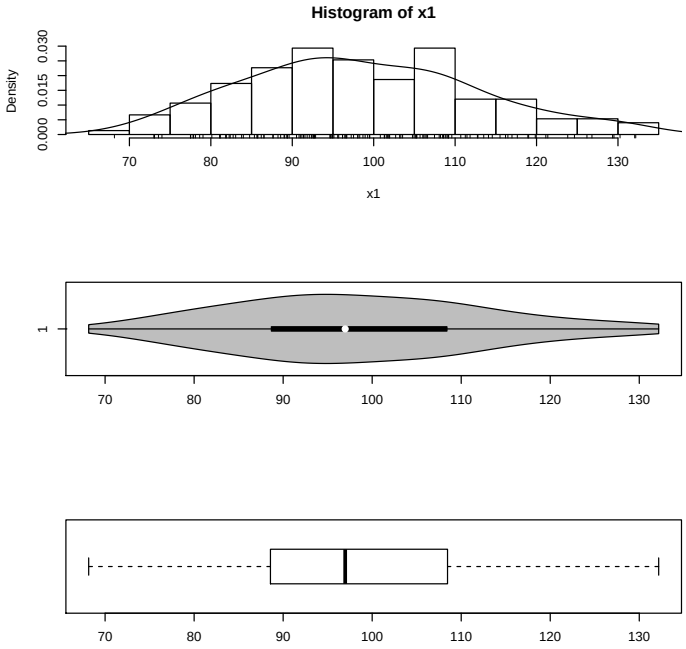
4.2 Testing Normality

An informal test of normality can be based on a **normal scores plot**, sometimes called a **rankit plot** or a **normal probability plot** or a **normal QQ plot** (QQ = quantile-quantile). You plot the quantiles of the data against the quantiles of the normal distribution, or **expected normal order statistics** (in a standard normal distribution) for a sample with the given number of observations. The normality assumption is plausible if the plot is fairly linear. I give below several plots often seen with real data, and what they indicate about the underlying distribution.

There are multiple ways to produce QQ plots in R. The shape can depend upon whether you plot the normal scores on the x-axis or the y-axis. It is conventional to plot the data on the y-axis and the normal scores on the x-axis.

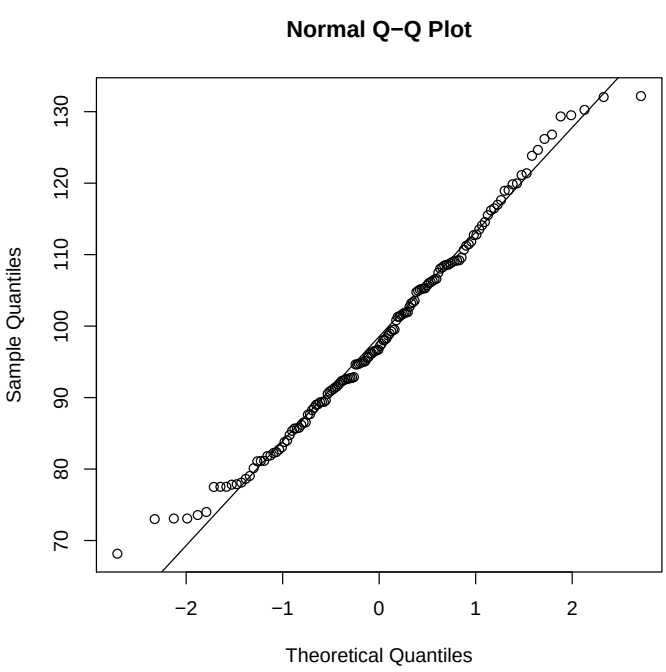
Let's start with some data from a normal distribution.

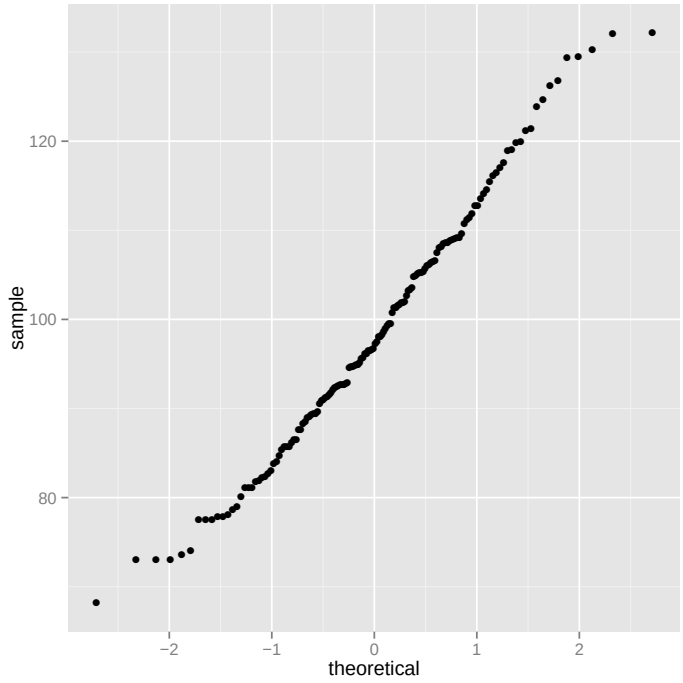
```
#### sample from normal distribution
x1 <- rnorm(150, mean = 100, sd = 15)

par(mfrow=c(3,1))
# Histogram overlaid with kernel density curve
hist(x1, freq = FALSE, breaks = 20)
points(density(x1), type = "l")
rug(x1)

# violin plot
library(vioplot)
vioplot(x1, horizontal=TRUE, col="gray")

# boxplot
boxplot(x1, horizontal=TRUE)
```



There are many ways to get adequate QQ plots. Consider how outliers shows up in the QQ plot. There may be isolated points on ends of the QQ plot, but only on the right side is there an outlier. How could you have identified that the right tail looks longer than the left tail from the QQ plot?

```
#### QQ plots
# R base graphics
par(mfrow=c(1,1))
# plots the data vs their normal scores
qqnorm(x1)
# plots the reference line
qqline(x1)


# ggplot2 graphics
library(ggplot2)
# http://had.co.nz/ggplot2/stat_qq.html
df <- data.frame(x1)
# stat_qq() below requires "sample" to be assigned a data.frame column
p <- ggplot(df, aes(sample = x1))
# plots the data vs their normal scores
p <- p + stat_qq()
print(p)
```





If you lay a straightedge along the bulk of the plot (putting in a regression line is not the right way to do it, even if it is easy), you see that the most extreme point on the right is a little below the line, and the last few points on the left a little above the line. What does this mean? The point on the right corresponds to a data value *more extreme* than expected from a normal distribution (the straight line is where expected and actual coincide). Extreme points on the right are above the line. What about the left? Extreme points there should be *above* the line — since the deviations from the line are above it on the left, those points are also *more extreme* than expected.

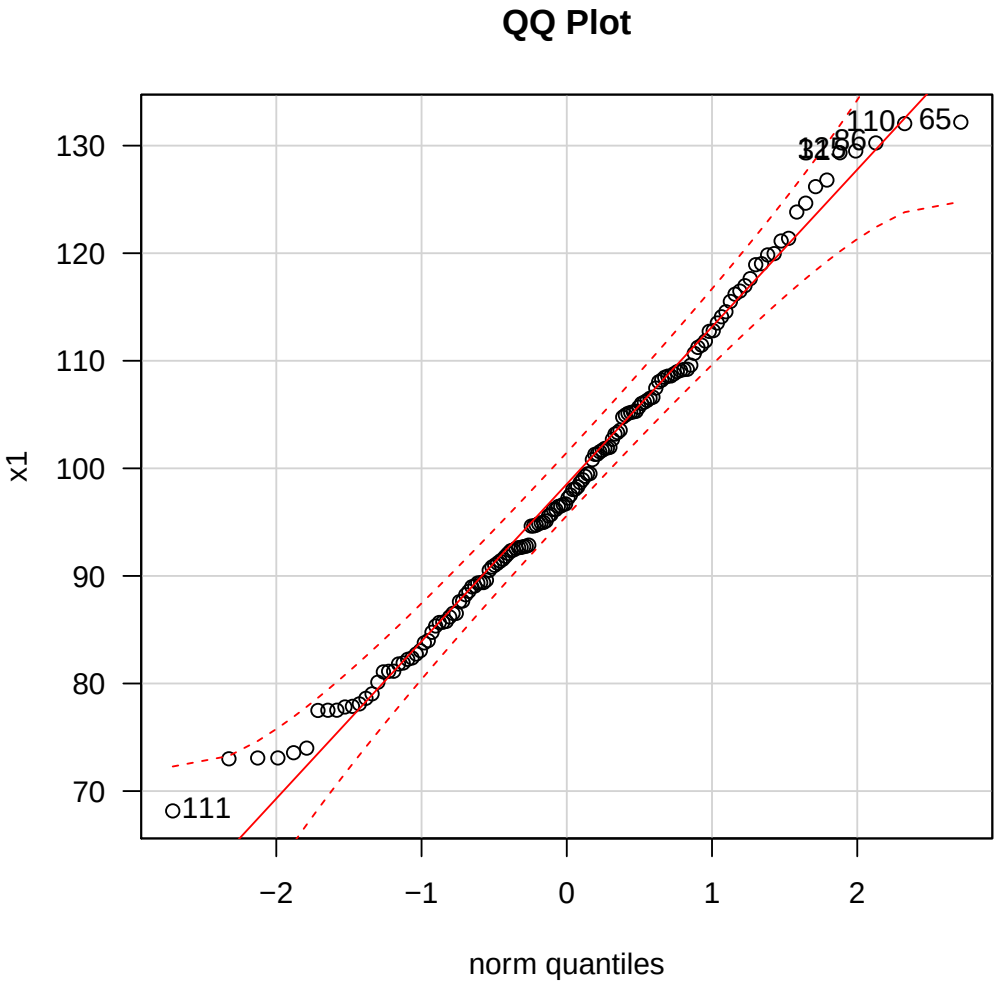
Even more useful is to add confidence intervals (point-wise, not family-wise — you will learn the meaning of those terms in the ANOVA section). You don't expect a sample from a normally distributed population to have a normal scores plot that falls exactly on the line, and the amount of deviation depends upon the sample size.

The best QQ plot I could find is available in the `car` package called `qqPlot`. Note that with the `dist=` option you can use this technique to see if the data appear from lots of possible distributions, not just normal.

```
par(mfrow=c(1,1))
# Normality of Residuals
library(car)
# qq plot for studentized resid
```

```
#   las    =  1 : turns labels on y-axis to read horizontally
#   id.n   =  n : labels n most extreme observations, and outputs to console
#   id.cex =  1 : is the size of those labels
#   lwd    =  1 : line width
qqPlot(x1, las = 1, id.n = 6, id.cex = 1, lwd = 1, main="QQ Plot")
```

```
##  65 110  86 125  31 111
## 150 149 148 147 146   1
```



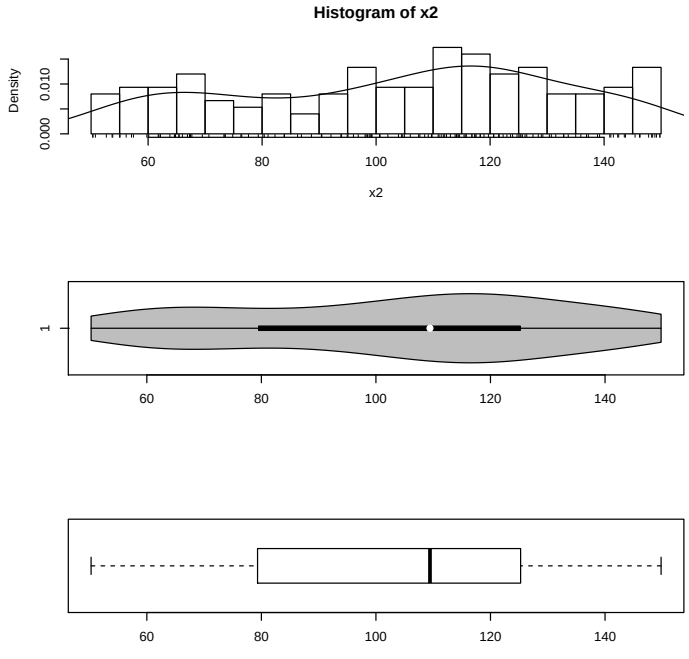
In this case the x-axis is labelled "norm quantiles". You only see a couple of data values outside the limits (in the tails, where it usually happens). You expect around 5% outside the limits, so there is no indication of non-normality here. I *did* sample from a normal population.

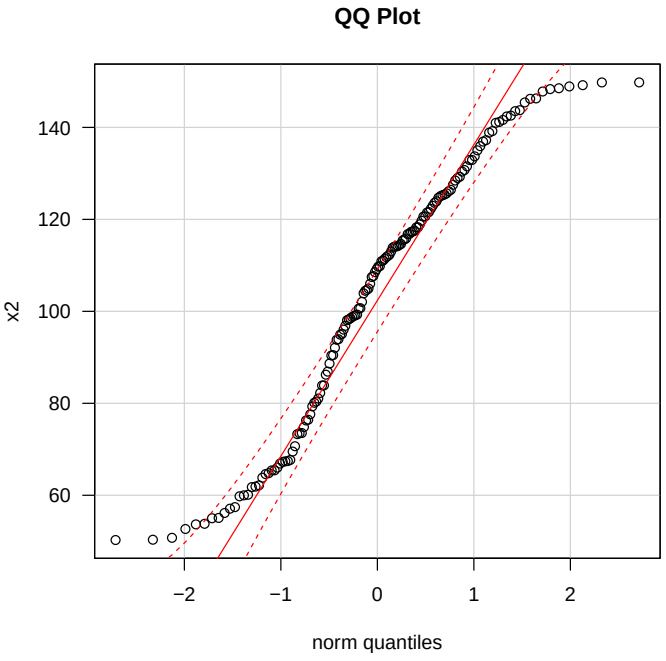
4.2.1 Normality tests on non-normal data

Let's turn to examples of sampling from other, non-normal distributions to see how the normal QQ plot identifies important features.

Light-tailed symmetric (Uniform)

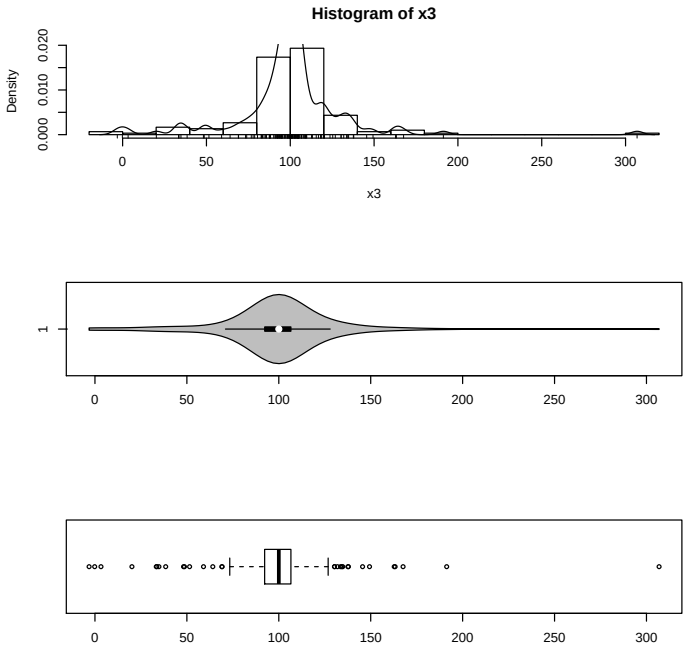
```
#### Light-tailed symmetric (Uniform)
# sample from uniform distribution
x2 <- runif(150, min = 50, max = 150)

par(mfrow=c(3,1))
# Histogram overlaid with kernel density curve
hist(x2, freq = FALSE, breaks = 20)
points(density(x2), type = "l")
rug(x2)

# violin plot
library(vioplot)
vioplot(x2, horizontal=TRUE, col="gray")

# boxplot
boxplot(x2, horizontal=TRUE)


par(mfrow=c(1,1))
qqPlot(x2, las = 1, id.n = 0, id.cex = 1, lwd = 1, main="QQ Plot")
```

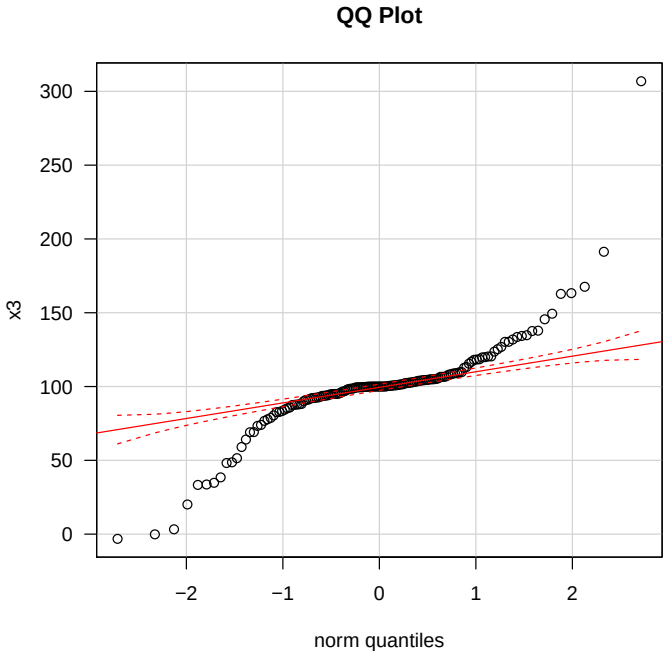




Heavy-tailed (fairly) symmetric (Normal-squared)

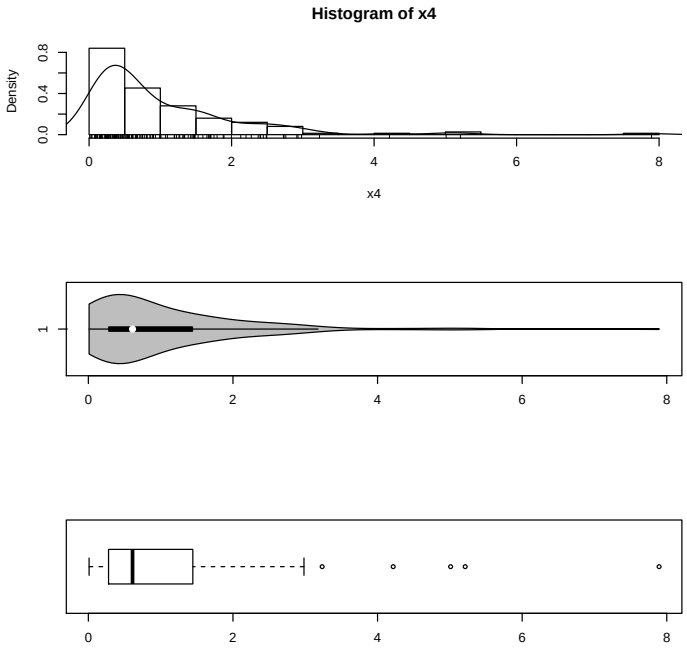
```
#### Heavy-tailed (fairly) symmetric (Normal-squared)
# sample from normal distribution
x3.temp <- rnorm(150, mean = 0, sd = 1)
x3 <- sign(x3.temp)*x3.temp^2 * 15 + 100

par(mfrow=c(3,1))
# Histogram overlaid with kernel density curve
hist(x3, freq = FALSE, breaks = 20)
points(density(x3), type = "l")
rug(x3)

# violin plot
library(vioplot)
vioplot(x3, horizontal=TRUE, col="gray")

# boxplot
boxplot(x3, horizontal=TRUE)


par(mfrow=c(1,1))
qqPlot(x3, las = 1, id.n = 0, id.cex = 1, lwd = 1, main="QQ Plot")
```

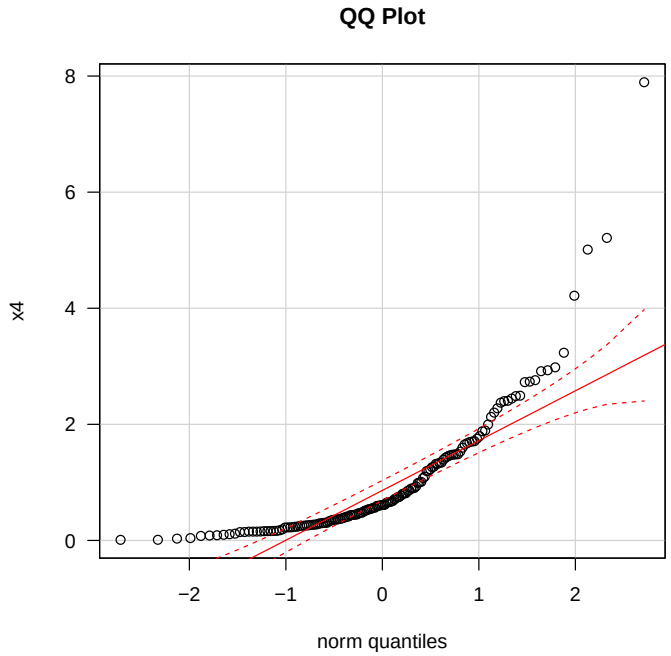




Right-skewed (Exponential)

```r
#### Right-skewed (Exponential)
# sample from exponential distribution
x4 <- rexp(150, rate = 1)

par(mfrow=c(3,1))
# Histogram overlaid with kernel density curve
hist(x4, freq = FALSE, breaks = 20)
points(density(x4), type = "l")
rug(x4)

# violin plot
library(vioplot)
vioplot(x4, horizontal=TRUE, col="gray")

# boxplot
boxplot(x4, horizontal=TRUE)


par(mfrow=c(1,1))
qqPlot(x4, las = 1, id.n = 0, id.cex = 1, lwd = 1, main="QQ Plot")
```

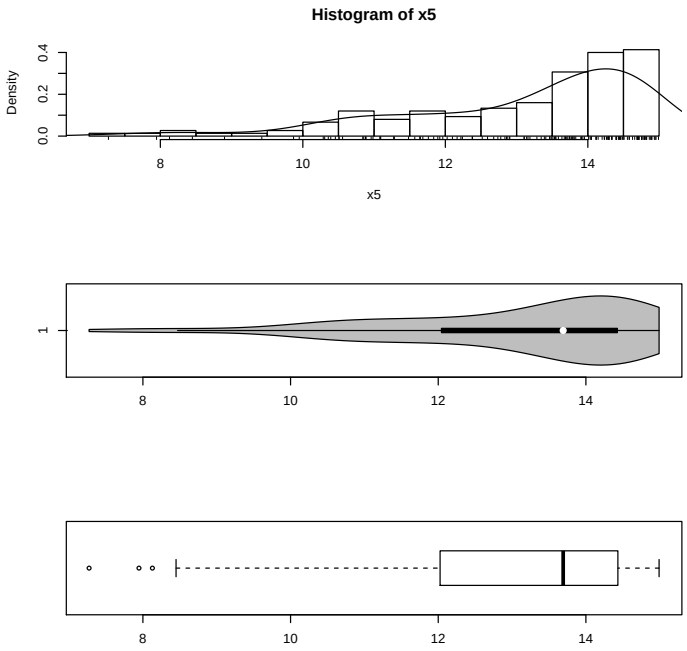


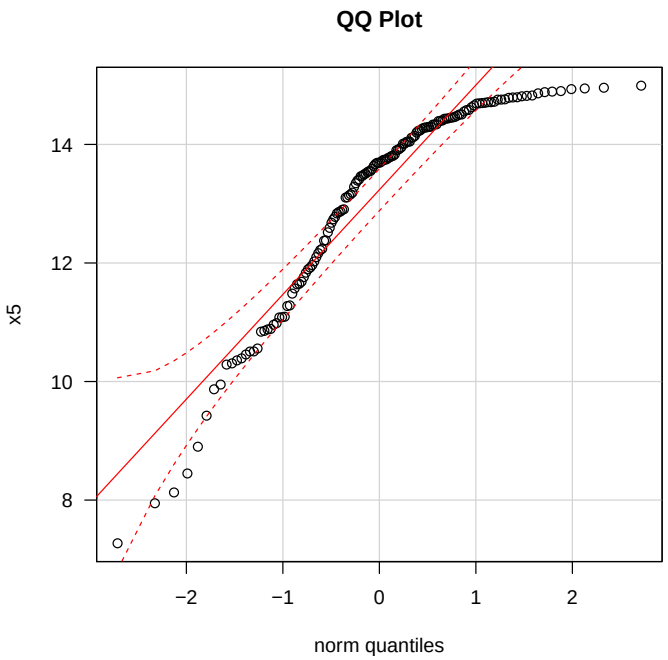


Left-skewed (Exponential, reversed)

```r
#### Left-skewed (Exponential, reversed)
# sample from exponential distribution
x5 <- 15 - rexp(150, rate = 0.5)
```

```r
par(mfrow=c(3,1))
# Histogram overlaid with kernel density curve
hist(x5, freq = FALSE, breaks = 20)
points(density(x5), type = "l")
rug(x5)

# violin plot
library(vioplot)
vioplot(x5, horizontal=TRUE, col="gray")

# boxplot
boxplot(x5, horizontal=TRUE)


par(mfrow=c(1,1))
qqPlot(x5, las = 1, id.n = 0, id.cex = 1, lwd = 1, main="QQ Plot")
```





Notice how striking is the lack of linearity in the QQ plot for all the non-normal distributions, particularly the symmetric light-tailed distribution where the boxplot looks fairly good. The QQ plot is a sensitive measure of normality. Let us summarize the patterns we see regarding tails in the plots:

	Tail	
Tail Weight	Left	Right
Light	Left side of plot points left	Right side of plot points right
Heavy	Left side of plot points down	Right side of plot points up

4.3 Formal Tests of Normality

A formal test of normality is based on the **correlation** between the data and the normal scores. The correlation is a measure of the strength of a linear relationship, with the sign of the correlation indicating the direction of the relationship (that is, + for increasing relationship, and − for decreasing). The correlation varies from −1 to +1. In a normal scores plot, you are looking for a correlation close to +1. Normality is rejected if the correlation is too small. Critical values for the correlation test of normality, which is commonly called the **Shapiro-Wilk** test, can be found in many texts.

R has several tests of normality. The **Shapiro-Wilk** test `shapiro.test()` is a base function. The R package `nortest` has five others: the Anderson-Darling test `ad.test()` is useful, related to the **Kolmogorov-Smirnov** (Lilliefors) test `lillie.test()` which is commonly used in many scientific disciplines but is essentially useless, the Cramer-von Mises test `cvm.test()`, and two more. Some packages also have the **Ryan-Joiner** test (closely related to the Shapiro-Wilk test).

Extreme outliers and skewness have the biggest effects on standard methods based on normality. The Shapiro-Wilk (SW) test is better at picking up these problems than the Kolmogorov-Smirnov (KS) test. The KS test tends to highlight deviations from normality in the center of the distribution. These types of deviations are rarely important because they do not have a noticeable effect on the operating characteristics of the standard methods. The AD and RJ tests are modifications designed to handle some of these objections.

Tests for normality may have low power in small to moderate sized samples. Visual assessment of normality is often more valuable than a formal test. The tests for the distributions of data above are below and in Figure 4.1.

Normal distribution

```
#### Formal Tests of Normality
shapiro.test(x1)
```

```
##
##  Shapiro-Wilk normality test
##
```

```
## data:  x1
## W = 0.98584, p-value = 0.1289
```

```
library(nortest)
ad.test(x1)
```

```
##
##  Anderson-Darling normality test
##
## data:  x1
## A = 0.40732, p-value = 0.3446
```

```
# lillie.test(x1)
cvm.test(x1)
```

```
##
##  Cramer-von Mises normality test
##
## data:  x1
## W = 0.05669, p-value = 0.4159
```

Light-tailed symmetric

```
shapiro.test(x2)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  x2
## W = 0.95252, p-value = 5.336e-05
```

```
library(nortest)
ad.test(x2)
```

```
##
##  Anderson-Darling normality test
##
## data:  x2
## A = 1.9426, p-value = 5.644e-05
```

```
# lillie.test(x2)
cvm.test(x2)
```

```
##
##  Cramer-von Mises normality test
##
## data:  x2
## W = 0.29567, p-value = 0.0003642
```

Heavy-tailed (fairly) symmetric

```
shapiro.test(x3)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  x3
## W = 0.79633, p-value = 3.587e-13
```

```
library(nortest)
ad.test(x3)
```

```
##
##  Anderson-Darling normality test
##
## data:  x3
## A = 9.1433, p-value < 2.2e-16
```

```
# lillie.test(x3)
cvm.test(x3)
```

```
## Warning in cvm.test(x3):  p-value is smaller
than 7.37e-10, cannot be computed more accurately
##
##  Cramer-von Mises normality test
##
## data:  x3
## W = 1.8248, p-value = 7.37e-10
```

Right-skewed

```
shapiro.test(x4)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  x4
## W = 0.74125, p-value = 5.872e-15
```

```
library(nortest)
ad.test(x4)
```

```
##
##  Anderson-Darling normality test
##
## data:  x4
## A = 9.3715, p-value < 2.2e-16
```

```
# lillie.test(x4)
cvm.test(x4)
```

```
## Warning in cvm.test(x4):  p-value is smaller
than 7.37e-10, cannot be computed more accurately
##
##  Cramer-von Mises normality test
##
## data:  x4
## W = 1.6537, p-value = 7.37e-10
```

Left-skewed

```
shapiro.test(x5)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  x5
## W = 0.8743, p-value = 5.933e-10
```

```
library(nortest)
ad.test(x5)
```

```
##
##  Anderson-Darling normality test
##
## data:  x5
## A = 6.0016, p-value = 7.938e-15
```

```
# lillie.test(x5)
cvm.test(x5)
```

```
##
##  Cramer-von Mises normality test
##
## data:  x5
## W = 1.0553, p-value = 9.648e-10
```

Figure 4.1: Normality tests for non-normal distributions

Example: Paired Differences on Sleep Remedies The following boxplot and normal scores plots suggest that the underlying distribution of differences (for the paired sleep data taken from the previous chapter) is reasonably symmetric, but heavy tailed. The p-value for the SW test of normality is 0.042, and for the AD test is 0.029, both of which call into question a normality assumption. A non-parametric test comparing the sleep remedies (one that does not assume normality) is probably more appropriate here. We will return to these data later.

```
# Normality tests
shapiro.test(sleep$d)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  sleep$d
## W = 0.83798, p-value = 0.04173
```

```
library(nortest)
ad.test(sleep$d)
```

```
##
##  Anderson-Darling normality test
##
## data:  sleep$d
## A = 0.77378, p-value = 0.02898
```

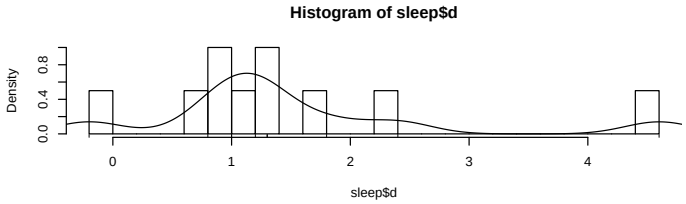
```
# lillie.test(sleep£d)
cvm.test(sleep$d)
```

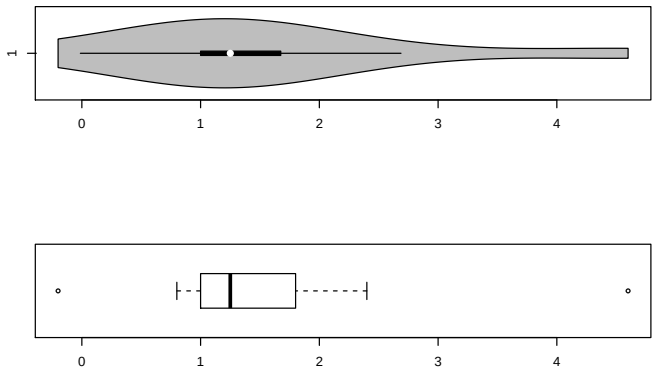
```
##
##  Cramer-von Mises normality test
##
## data:  sleep$d
## W = 0.13817, p-value = 0.02769
```

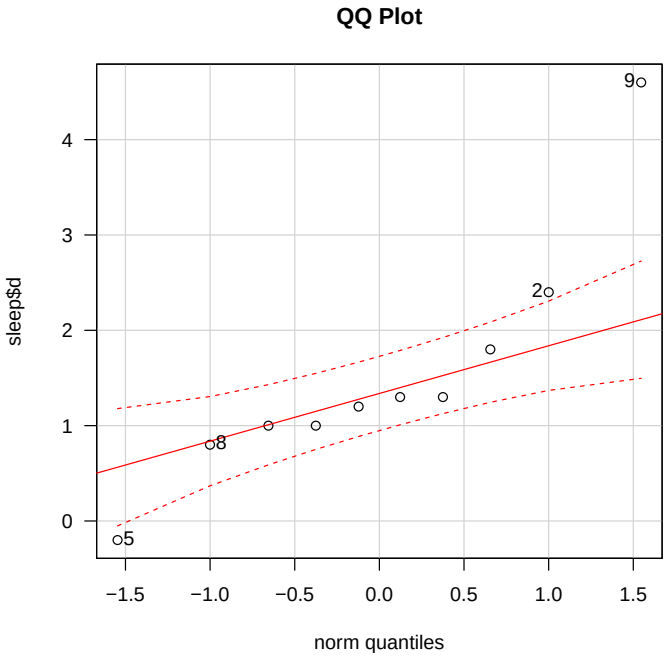
```
# plot of data
par(mfrow=c(3,1))
# Histogram overlaid with kernel density curve
hist(sleep$d, freq = FALSE, breaks = 20)
points(density(sleep$d), type = "l")
rug(sleep$d)

# violin plot
library(vioplot)
vioplot(sleep$d, horizontal=TRUE, col="gray")

# boxplot
boxplot(sleep$d, horizontal=TRUE)
```

```
# QQ plot
par(mfrow=c(1,1))
qqPlot(sleep$d, las = 1, id.n = 4, id.cex = 1, lwd = 1, main="QQ Plot")
```

```
##  9  5  2  8
## 10  1  9  2
```







Example: Androstenedione Levels This is an independent two-sample problem, so you must look at normal scores plots for males and females.

The AD test p-value and the SW test p-value for testing normality exceeds 0.10 in each sample. Thus, given the sample sizes (14 for men, 18 for women), we have insufficient evidence (at $\alpha = 0.05$) to reject normality in either population.

The women's boxplot contains two mild outliers, which is highly unusual when sampling from a normal distribution. The tests are possibly not powerful enough to pick up this type of deviation from normality in such a small sample. In practice, this may not be a big concern. The two mild outliers probably have a small effect on inferences in the sense that non-parametric methods would probably lead to similar conclusions here.

```
library(ggplot2)
p1 <- ggplot(andro, aes(x = sex, y = level, fill=sex))
p1 <- p1 + geom_boxplot()
```

Men

```
shapiro.test(men)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  men
## W = 0.90595, p-value = 0.1376
```

```
library(nortest)
ad.test(men)
```

```
##
##  Anderson-Darling normality test
##
## data:  men
## A = 0.4718, p-value = 0.2058
```

```
# lillie.test(men)
cvm.test(men)
```

```
##
##  Cramer-von Mises normality test
##
## data:  men
## W = 0.063063, p-value = 0.3221
```

Women

```
shapiro.test(women)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  women
## W = 0.95975, p-value = 0.5969
```

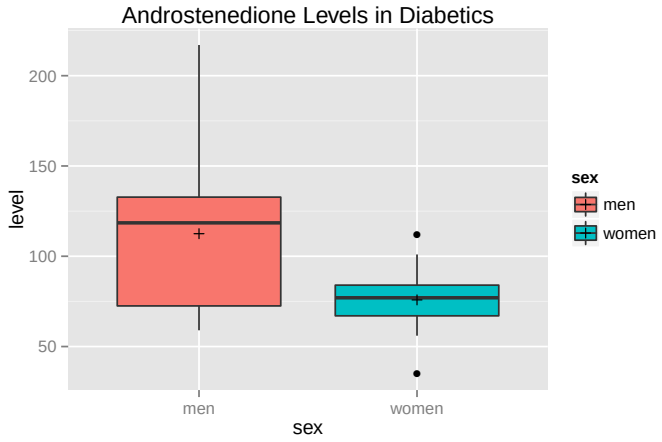
```
library(nortest)
ad.test(women)
```

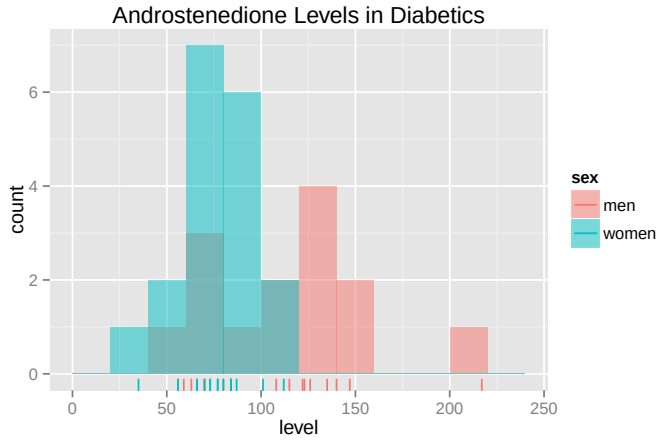
```
##
##  Anderson-Darling normality test
##
## data:  women
## A = 0.39468, p-value = 0.3364
```

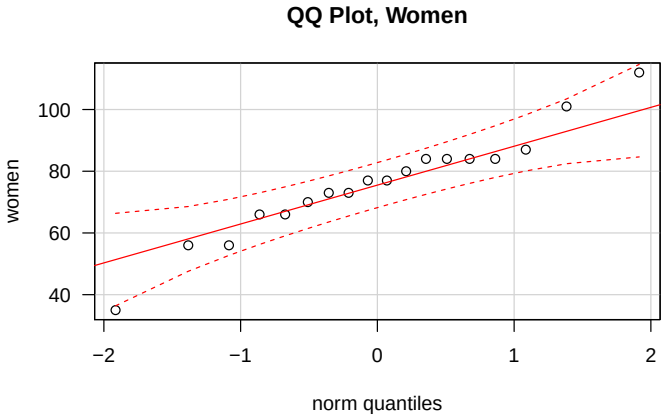
```
# lillie.test(women)
cvm.test(women)
```

```
##
##  Cramer-von Mises normality test
##
## data:  women
## W = 0.065242, p-value = 0.3057
```

```
# add a "+" at the mean
p1 <- p1 + stat_summary(fun.y = mean, geom = "point", shape = 3, size = 2)
#p1 <- p1 + coord_flip()
p1 <- p1 + labs(title = "Androstenedione Levels in Diabetics")
#print(p1)

p2 <- ggplot(andro, aes(x = level, fill=sex))
p2 <- p2 + geom_histogram(binwidth = 20, alpha = 0.5, position="identity")
p2 <- p2 + geom_rug(aes(colour=sex))
p2 <- p2 + labs(title = "Androstenedione Levels in Diabetics")
#print(p2)

library(gridExtra)
grid.arrange(p1, p2, ncol=1)


# QQ plot
par(mfrow=c(2,1))
qqPlot(men, las = 1, id.n = 0, id.cex = 1, lwd = 1, main="QQ Plot, Men")
qqPlot(women, las = 1, id.n = 0, id.cex = 1, lwd = 1, main="QQ Plot, Women")
```



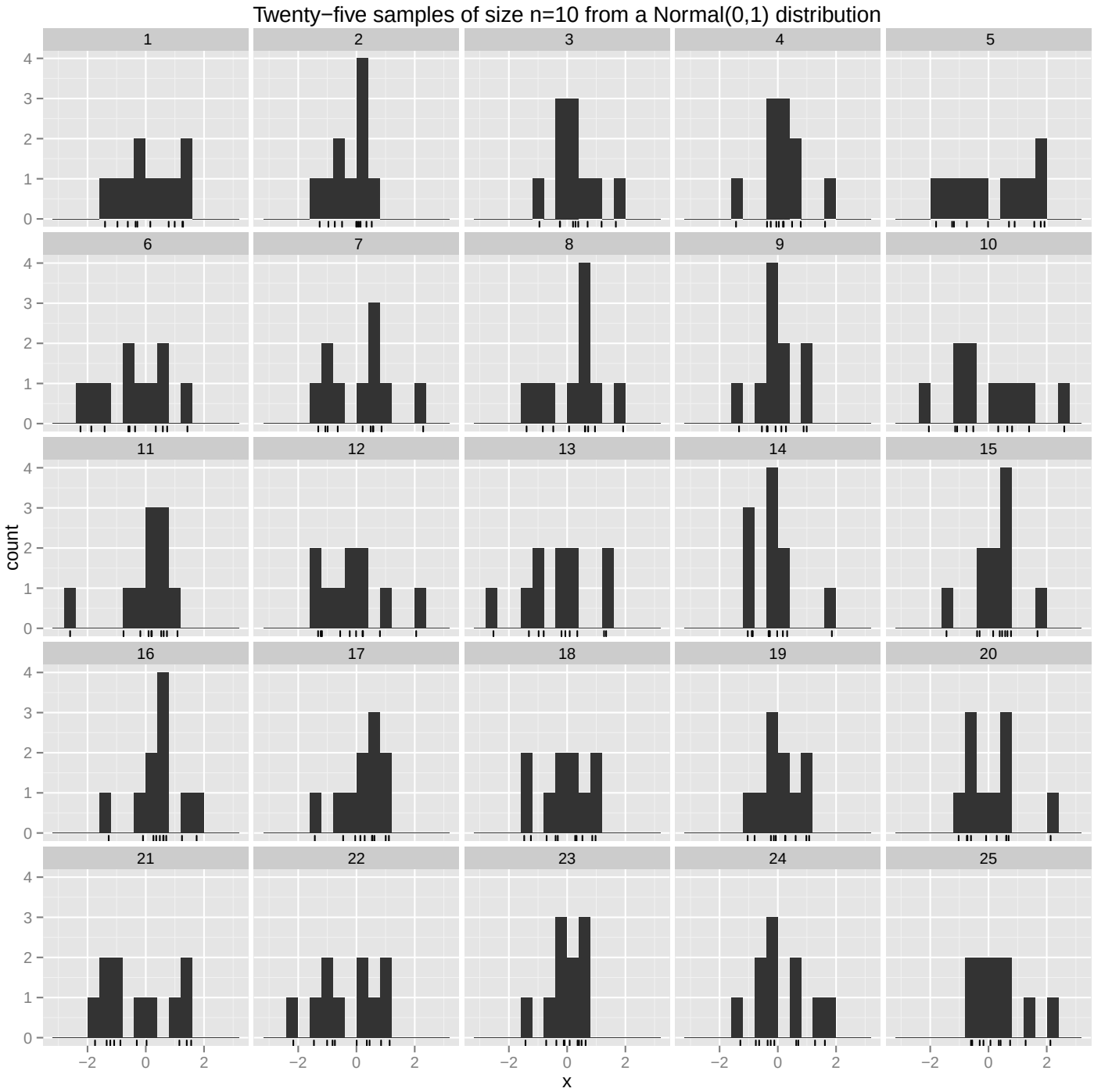




Most statisticians use graphical methods (boxplot, normal scores plot) to assess normality, and do not carry out formal tests.

You may be surprised at how variable 10 observations from a Normal(0,1) distribution looks like; here are 25 samples.

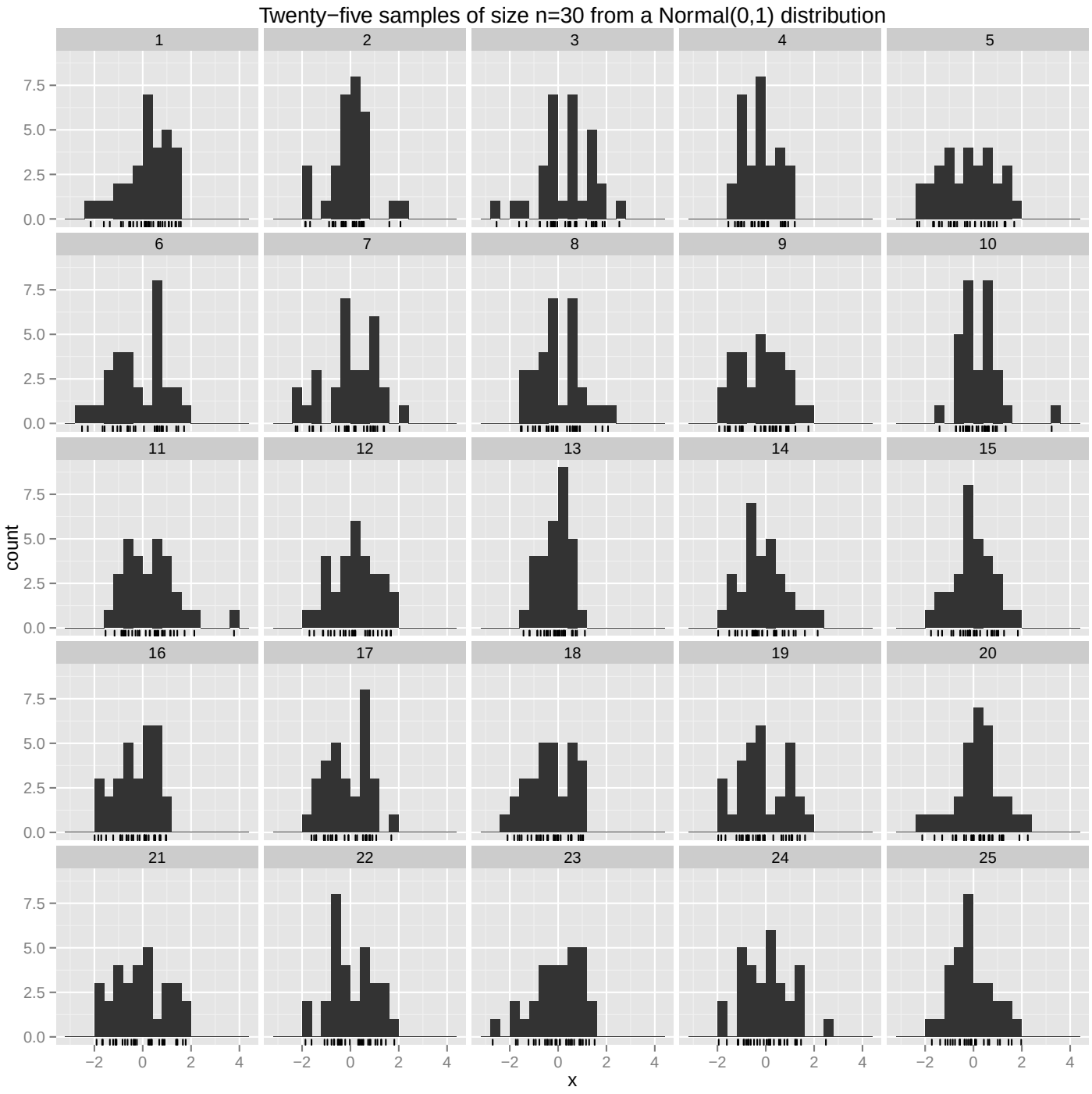
```
n = 10
r = 5
norm.many <- data.frame(id = rep(seq(1:r^2), n)
                        , x = rnorm(r^2 * n)
                         )
library(ggplot2)
p <- ggplot(norm.many, aes(x = x))
p <- p + geom_histogram(binwidth = 0.4)
p <- p + geom_rug()
p <- p + facet_wrap(~ id, ncol = r)
p <- p + labs(title = "Twenty-five samples of size n=10 from a Normal(0,1) distribution")
print(p)
```



. . . and here are samples of size $n = 30$.

```
n = 30
r = 5
norm.many <- data.frame(id = rep(seq(1:r^2), n)
                        , x = rnorm(r^2 * n)
                         )
library(ggplot2)
p <- ggplot(norm.many, aes(x = x))
p <- p + geom_histogram(binwidth = 0.4)
p <- p + geom_rug()
```

```
p <- p + facet_wrap(~ id, ncol = r)
p <- p + labs(title = "Twenty-five samples of size n=30 from a Normal(0,1) distribution")
print(p)
```



By viewing many versions of this of varying samples sizes you'll develop your intuition about what a normal sample looks like.

4.4 Testing Equal Population Variances

In the independent two sample t-test, some researchers test $H_0 : \sigma_1^2 = \sigma_2^2$ as a means to decide between using the pooled-variance procedure or Satterthwaite's methods. They suggest the pooled t-test and CI if H_0 is not rejected, and Satterthwaite's methods otherwise.

There are a number of well-known tests for equal population variances, of which Bartlett's test and Levene's test are probably the best known. Bartlett's test assumes the population distributions are normal, and is the best test when this is true. In practice, unequal variances and non-normality often go hand-in-hand, so you should check normality prior to using Bartlett's test. It is sensitive to data which is not non-normally distributed, thus it is more likely to return a "false positive" (reject H_0 of equal variances) when the data is non-normal. Levene's test is more robust to departures from normality than Bartlett's test; it is in the `car` package. Fligner-Killeen test is a non-parametric test which is very robust against departures from normality.

I will now define **Bartlett's test**, which assumes normally distributed data. As above, let $n^* = n_1 + n_2 + \cdots + n_k$, where the n_is are the sample sizes from the k groups, and define

$$v = 1 + \frac{1}{3(k-1)} \left(\sum_{i=1}^{k} \frac{1}{n_i - 1} - \frac{1}{n^* - k} \right).$$

Bartlett's statistic for testing $H_0 : \sigma_1^2 = \cdots = \sigma_k^2$ is given by

$$B_{obs} = \frac{2.303}{v} \left\{ (n-k) \log(s^2_{\text{pooled}}) - \sum_{i=1}^{k} (n_i - 1) \log(s_i^2) \right\},$$

where s^2_{pooled} is the pooled estimator of variance and s_i^2 is the estimated variance based on the i^{th} sample.

Large values of B_{obs} suggest that the population variances are unequal. For a size α test, we reject H_0 if $B_{\text{obs}} \geq \chi^2_{k-1,\text{crit}}$, where $\chi^2_{k-1,\text{crit}}$ is the upper-α percentile for the χ^2_{k-1} (chi-squared) probability distribution with $k-1$ degrees

of freedom. A generic plot of the χ^2 distribution is given below. A p-value for the test is given by the area under the chi-squared curve to the right of B_{obs}.

Example: Androstenedione Levels The sample standard deviations and samples sizes are: $s_1 = 42.8$ and $n_1 = 14$ for men and $s_2 = 17.2$ and $n_2 = 18$ for women. The sample standard deviations appear to be very different, so I would not be surprised if the test of equal population variances is highly significant. The output below confirms this: the p-values for Bartlett's F-test, Levene's Test, and Fligner-Killeen test are all much smaller than 0.05. An implication is that the standard pooled-CI and test on the population means is inappropriate.

```
#### Testing Equal Population Variances
# numerical summaries
c(mean(men), mean(women), sd(men), sd(women))
## [1] 112.50000  75.83333  42.75467  17.23625
c(IQR(men), IQR(women), length(men), length(women))
## [1] 60.25 17.00 14.00 18.00
## Test equal variance
# assumes populations are normal
bartlett.test(level ~ sex, data = andro)
##
##  Bartlett test of homogeneity of variances
##
## data:  level by sex
## Bartlett's K-squared = 11.199, df = 1, p-value = 0.0008183
# does not assume normality, requires car package
library(car)
leveneTest(level ~ sex, data = andro)
## Levene's Test for Homogeneity of Variance (center = median)
##       Df F value  Pr(>F)
## group  1  7.2015 0.01174 *
##       30
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
# nonparametric test
fligner.test(level ~ sex, data = andro)
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  level by sex
```

```
## Fligner-Killeen:med chi-squared = 5.8917, df = 1, p-value =
## 0.01521
```

4.5 Small sample sizes, a comment

In Daniel Kahneman's "Thinking, Fast and Slow" (Ch 10), he discusses "The Law of Small Numbers" (in contrast to the Law of Large Numbers). As an example from statisticians Howard Wainer and Harris Zwerling, he makes this observation about the incidence of kidney cancer in the 3,141 counties of the United States. "The counties in which the incidence of kidney cancer is *lowest* are mostly rural, sparsely populated, and located in traditionally Republican states in the Midwest, the South, and the West. What do you make of this?" The statistians comment: "It is both easy and tempting to infer that their low cancer rates are directly due to the clean living of the rural lifestyle — no air pollution, no water pollution, access to fresh food without additives." This makes perfect sense.

"Now consider the counties in which the incidence of kidney cancer is highest. These ailing counties tend to be mostly rural, sparsely populated, and located in traditionally Republican states in the Midwest, the South, and the West." Tongue-in-cheek, Wainer and Zwerling comment: "It is easy to infer that their high cancer rates might be directly due to the poverty of the rural lifestyle — no access to good medical care, a high-fat diet, and too much alcohol, too much tobacco." Something is wrong, of course. The rural lifestyle cannot explain both very high and very low incidence of kidney cancer.

The key factor is not that the counties were rural or predominantly Republican. It is that rural counties have small populations. The law of large numbers says that as sample sizes increase that the sample statistic converges to the population proportion, that is, large samples are more precise than small samples. What Kahneman is calling the law of small numbers warns that small samples yield extreme results more often than large samples do.