Statistics with R

Hypothesis testing and distributions

Steven Buechler

Department of Mathematics
276B Hurley Hall; 1-6233

Fall, 2007

How do we study vectors?

Readings across a population

How do we find patterns in accumulated readings of a single variable?

- For a categorical variable, what is the distribution of values?
- If it's a continuous variable is it normally distributed?
- Are there phenotypic subsets where the distributions are different? How can we test this?

How do we study vectors?

Readings across a population

How do we find patterns in accumulated readings of a single variable?

- For a categorical variable, what is the distribution of values?
- If it's a continuous variable is it normally distributed?
- Are there phenotypic subsets where the distributions are different? How can we test this?

How do we study vectors?

Readings across a population

How do we find patterns in accumulated readings of a single variable?

- For a categorical variable, what is the distribution of values?
- If it's a continuous variable is it normally distributed?
- Are there phenotypic subsets where the distributions are different? How can we test this?

Examining Categorical Data

With categorical data about all you can do is examine the frequency table:

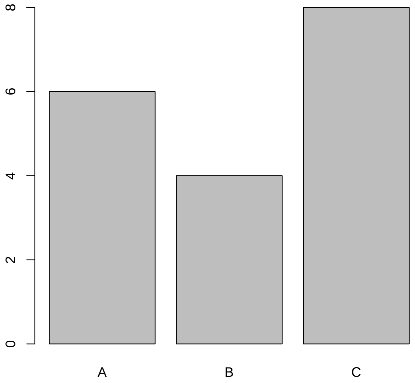
```
> f <- factor(c(rep("A", 6), rep("B", 4),
+     rep("C", 8)))
> table(f)
```

```
f
A B C
6 4 8
```

Visualizing Categorical Data

However, when the levels are numerous it may be helpful to produce a barplot.

```
> plot(f)
```



The Plot Function

- `plot` can create a wide variety of graphics depending on the input and user-defined parameters. Options allow on the fly visualization with one-line commands, or publication-quality annotated diagrams.
- `plot(...)` creates a graphics panel displaying the result. Good for simply picturing what you're doing. These can be saved for future use.
- Other commands can generate JPEG, PNG, PDF, etc., files containing your plots that can be included in a Word document.

The Plot Function

- `plot` can create a wide variety of graphics depending on the input and user-defined parameters. Options allow on the fly visualization with one-line commands, or publication-quality annotated diagrams.
- `plot(...)` creates a graphics panel displaying the result. Good for simply picturing what you're doing. These can be saved for future use.
- Other commands can generate JPEG, PNG, PDF, etc., files containing your plots that can be included in a Word document.

The Plot Function

- `plot` can create a wide variety of graphics depending on the input and user-defined parameters. Options allow on the fly visualization with one-line commands, or publication-quality annotated diagrams.
- `plot(...)` creates a graphics panel displaying the result. Good for simply picturing what you're doing. These can be saved for future use.
- Other commands can generate JPEG, PNG, PDF, etc., files containing your plots that can be included in a Word document.

Values of a Continuous Variable

Readings from a continuous variable may need much more analysis to understand.

Given: Vectors of readings, `x1, x2, x3`, perhaps by different experimenters. E.g.,

```
> x1

 [1] 1.4446 1.5745 1.5533 1.0095 0.9227 1.6656
 [7] 1.1701 1.6607 0.9651 0.9778 0.6729 1.4089
[13] 1.8254 0.5351 1.8276 1.3317 2.0544 1.3893
[19] 1.2932 1.2404
```

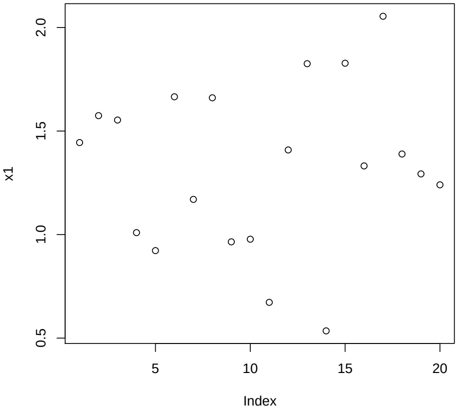
First visualization is a scatter plot that plots the index on the x axis and the value on the y axis. This is the default when `plot` is given a numeric vector.

Scatter Plots

For viewing the distribution of values

Create a scatter plot of `x1`:

```
> plot(x1)
```



Scatter Plots - Refined

Can add more points and lines

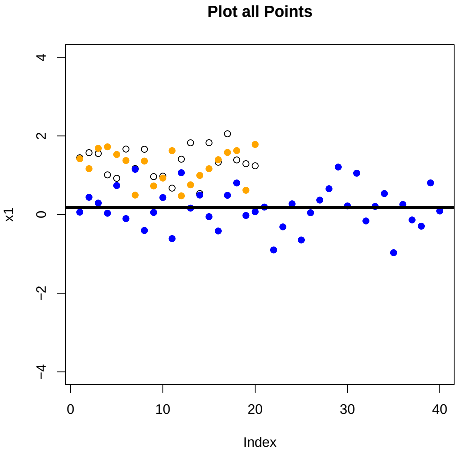
Create a scatter plot of x1, x2, x3 with notation that separates them.

```
> plot(x1, xlim = c(1, 40), ylim = c(-4,
+     4), main = "Plot all Points")
> points(x2, pch = 19, col = "orange")
> points(x3, pch = 19, col = "blue")
> abline(h = mean(x3), lwd = 3)
```

Scatter Plots - Refined

Can add more points and lines

Create a scatter plot of `x1, x2, x3` with notation that separates them.

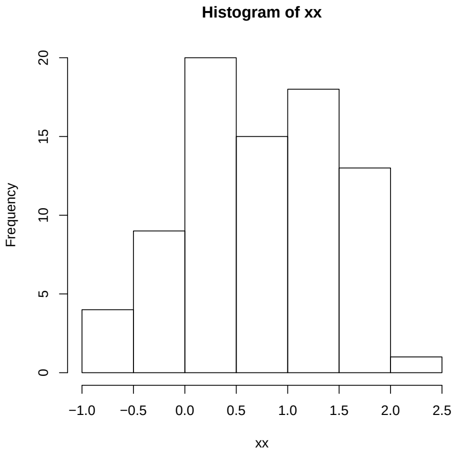


Histograms for Frequency

How many values lie in a “bin”?

Combine all points and create a historgam.

```
> xx <- c(x1, x2, x3)
> hist(xx)
```



Histograms in R

There is a theory of histograms that suggest a bin width that is most informative. As expected it is possible to fill or cross-hatch the bars, overlay other plots, etc.

When is a Variable Normally Distributed?

There are several levels of precision in answering this. In deciding if a particular statistical test can be applied it is usually good enough if it *looks* normal. However, histograms or the related density plots aren't the accepted way.

What is a Distribution?

Reminder of Basic Definitions

A discrete random variable is a numerical quantity that takes values with some randomness from a discrete set; often a subset of integers. The probability distribution of a discrete random variable specifies the probability associated with each possible value.

What is a Distribution?

Continuous Random Variables

A continuous random variable X takes values in an interval of real numbers. There is a probability associated with X falling between two numbers $a < b$. The density function $f_X(x)$ is such that $Prob(a \leq X \leq b)$ is the area bounded by the graph of $y = f_X(x)$, the $x-$axis, and the vertical lines $x = a$ and $x = b$. In other words, $Prob(a \leq X \leq b) = \int_a^b f_X(x)dx$.

Distribution Functions and Quantiles

For Continuous Random Variables

The cumulative distribution function for X is $F_X(x) = Prob(X \leq x)$. It's the area under the curve of the density function to the left of x. The quantile function of X, $Q_X(u)$ is the value x of X such that $Prob(X \leq x) = u$. In other words, it is the inverse of $F_X(x)$.

Normal Distribution

A normally distributed random variable with mean μ and standard deviation σ is one with density function

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

It's graph is a bell curve centered at μ whose "spread" is determined by σ. The standard normal distribution is one with mean 0 and standard deviation 1.

Normal Distribution

Computing Values in *R*

The distribution function for the normal with mean = `'mean'` and standard deviation = `'sd'` is `pnorm(x, mean, sd)`. The quantile function of the normal is `qnorm(p, mean, sd)`. The function `rnorm(n, mean, sd)` randomly generates `n` values of a normally distributed random variable with given mean and sd. Default values: `mean=0, sd=1`.

Normal Distribution

Examples

```
> pnorm(2)
[1] 0.9772
> pnorm(0)
[1] 0.5
> qnorm(0.95)
[1] 1.645
> qnorm(0.025, mean = 2, sd = 0.5)
[1] 1.02
> rnorm(4, 2, 2)
[1]  1.00381  3.94258 -0.01914  2.30742
```

Data Are Finite Samples from X

Collected data are finitely many values from a random variable X; a finite sample. While X has a precise distribution we can only estimate it from the finite sample.

How do we test a sample against a hypothesized distribution? Is it normal, t, Chi-squared, . . . ? More generally, how do we test when two samples come from the same distribution?

Quantile-Quantile Plots

Quantiles are Calculable

It's hard going from a sample to estimate a density function. We can, however, calculate the sample quantiles. If two samples come from the same distribution they should have the same quantiles. Given two samples A and B let x_i and y_i be the i^{th} quantile of A and B, resp. Graph the points (x_i, y_i) to form the Quantile-Quantile Plot (Q-Q plot) of A and B.

When A and B have the same distribution the Q-Q plot is a 45^o straight line.

Q-Q Normal Plots

In many instances we need to know if a sample comes from a normal distribution. In this case we compare the sample quantiles against the calculated quantiles of a normal distribution.

Q-Q Normal Plots

The Q-Q Normal Plot of A is then the Q-Q plot of A against the standard normal distribution.

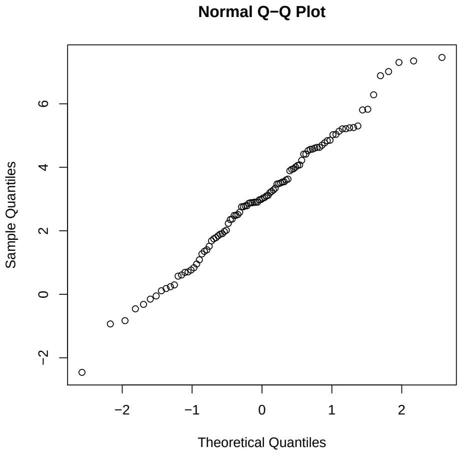
NOTE: This will be a straight line if the distribution of A is normal of any mean and standard deviation.

Happily, there is an R function that does all of this: `qqnorm`.

Q-Q Normal Plots

Examples

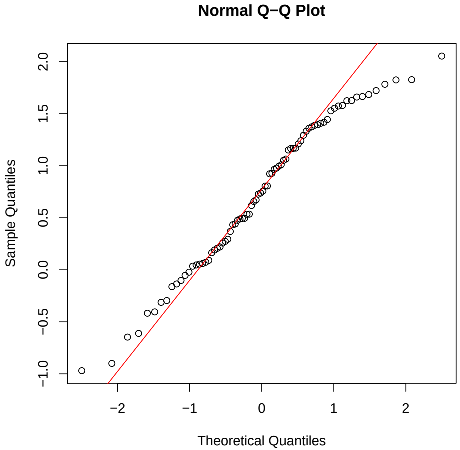
```
> v <- rnorm(100, 3, 2)
> qqnorm(v)
```



Q-Q Normal Plots of Examples

Recall that `xx` contains all 80 points.

```
> qqnorm(xx)
> qqline(xx, col = "red")
```



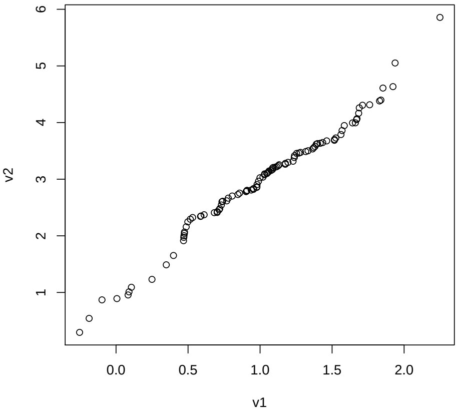
Q-Q Line

The Q-Q line is drawn so that it passes through the first and third quantile. All of the points should be on this line when the sample is normal.

In this example the distribution appears to be shifted to the left from a normal.

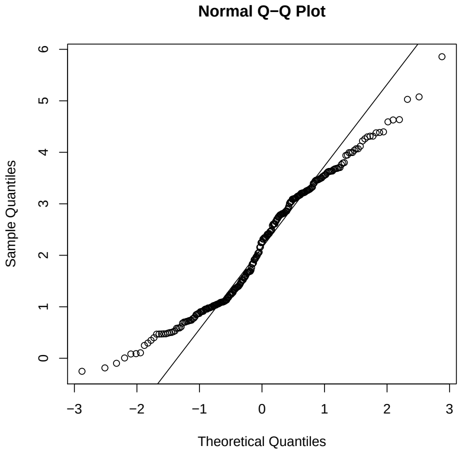
Q-Q Plot to Compare Two Samples

```
> v1 <- rnorm(100, 1, 0.5)
> v2 <- rnorm(150, 3, 1)
> qqplot(v1, v2)
```



Q-Q Plot to Compare Two Samples

```
> qqnorm(c(v1, v2))
> qqline(c(v1, v2))
```



Classical Hypothesis Testing

Review or Reading Assignment

Test of a null hypothesis against an alternative hypothesis. There are five steps, the first four of which should be done before inspecting the data.

Step 1. Declare the null hypothesis H_0 and the alternative hypothesis H_1.

In a sequence matching problem H_0 may be that two sequences are uniformly independent, in which case the probability of a match is 0.25. H_1 may be "probability of a match $= 0.35$", or "probability of a match > 0.25".

Classical Hypothesis Testing

Types of hypotheses

A hypothesis that completely specifies the parameters is called simple. If it leaves some parameter undetermined it is composite. A hypothesis is one-sided if it proposes that a parameter is $>$ some value or $<$ some value; it is two-sided if it simply says the parameter is $\neq$ some value.

Types of Error

Rejecting H_0 when it is actually true is called a Type I Error. In biomedical settings it can be considered a false positive. (Null hypothesis says “nothing is happening” but we decide “there is disease”.)

Step 2. Specify an acceptable level of Type I error, α, normally 0.05 or 0.01.

This is the threshold used in deciding to reject H_0 or not. If $\alpha = 0.05$ and we determine the probability of our data assuming H_0 is 0.0001, then we reject H_0.

The Test Statistic

Step 3. Select a test statistic.

This is a quantity calculated from the data whose value leads me to reject the null hypothesis or not. For matching sequences one choice would be the number of matches. For a contingency table compute Chi-squared. Normally compute the value of the statistic from the data assuming H_0 is true.

A great deal of theory, experience and care can go into selecting the right statistic.

The Critical Value or Region

Step 4. Identify the values of the test statistic that lead to rejection of the null hypothesis.

This is a quantity calculated from the data whose value leads me to reject the null hypothesis or not. For matching sequences one choice would be the number of matches. For a contingency table compute Chi-squared. Normally compute the value of the statistic from the data assuming H_0 is true. *A great deal of theory, experience and care can go into selecting the right statistic.*

The Critical Value or Region

Example

The statistic for the number Y of matches between two sequences of nucleotides is a binomial random variable. Let n be the lengths of the two sequences (assume they are the same). Under the null hypothesis that there are only random connections between the sequences the probability of a match at any point is $p = 0.25$. We reject the null hypothesis if the observed value of Y is so large that the chance of obtaining it is < 0.05.

The Critical Value or Region

Example

There is a specific formula for the probability of Y matches in n "trials" with probability of a match $= 0.25$. We can similarly calculate the significance threshold K so that

$$Prob(Y \geq K|p = 0.25) = 0.05.$$

When $n = 100$, $Prob(Y \geq 32) = .069$ and $Prob(Y \geq 33) = .044$. Take as the significance threshold 33. Reject the null hypothesis if there are at least 33 matches.

Obtain the Data and Execute

Step 5. Obtain the data, calculate the value of the statistic assuming the null hypothesis and compare with the threshold.

P-Values

Substitute for Step 4

Once the data are obtained calculate the null hypothesis probability of obtaining the observed value of the statistic or one more extreme. This is called the p-value. If it is $<$ the selected Type I Error threshold then we reject the null hypothesis.

P-Values

Example

Compare sequences of length 26 under the null hypothesis of only random matches; i.e., $p = 0.25$. Suppose there are 11 matches in our data. In a binomial distribution of length 26 with $p = 0.25$ the probability of ≥ 11 matches is about 0.04. So, with the Type I Error rate, α, at 0.05 we would reject the null hypothesis.