MIE237

Neil Montgomery
2016-01-22

A taste of categorical data 9.10, 10.13

Example "gas pipeline data"

Leak	Size	Material	Pressure
No	1.00	Aldyl A	Low
No	1.00	Steel	Med
No	1.50	Aldyl A	Low
Yes	1.50	Aldyl A	Low
No	1.50	Steel	Med
Yes	1.75	Aldyl A	Low
Yes	1.00	Aldyl A	Med
...	...	...	...

Numbers of interest

- Counts and *proportions* of one- and multi-way classifications.
- One-way on `Leak`:

```
## Source: local data frame [2 x 3]
##
##    Leak Count Proportion
##   (chr) (int)      (dbl)
## 1    No   802      0.802
## 2   Yes   198      0.198
```

- Note: avoid "percentages", which are really just a way of formatting proportions for human visual consumption.

numbers…two-way classification

- A few different styles…

```
##      Size
## Leak    1 1.5 1.75
##   No  329 249  224
##   Yes  93  52   53
```

```
## Source: local data frame [6 x 3]
## Groups: Leak [?]
##
##    Leak  Size     n
##   (chr) (dbl) (int)
## 1    No  1.00   329
## 2    No  1.50   249
## 3    No  1.75   224
## 4   Yes  1.00    93
## 5   Yes  1.50    52
## 6   Yes  1.75    53
```

numbers…two-way classification

- Adding marginal totals:

```
##      Size
## Leak     1  1.5 1.75  Sum
##   No   329  249  224  802
##   Yes   93   52   53  198
##   Sum  422  301  277 1000
```

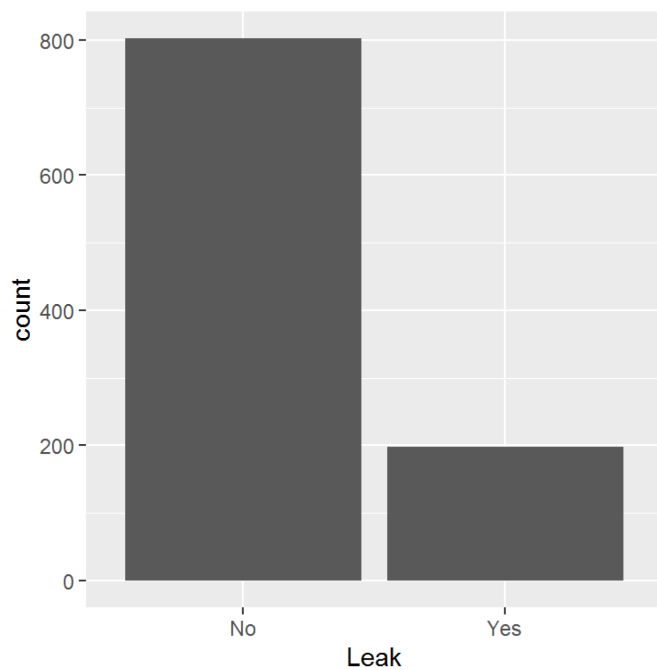
- Proportions rather than counts:

```
##      Size
## Leak      1   1.5  1.75   Sum
##   No  0.329 0.249 0.224 0.802
##   Yes 0.093 0.052 0.053 0.198
##   Sum 0.422 0.301 0.277 1.000
```

graphical summaries

- Pretty much "bar plot" or "bar chart" and friends

```
pipeline %>%
  ggplot(aes(x = Leak)) + geom_bar()
```



R diversion: factor

- In R a `factor` is a special kind of variable, specifically for categorical variables. The values of a `factor` variable are restricted to certain `levels`.
- Let's look at what R thinks the variables of `pipeline` are made of:

```
str(pipeline)
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    1000 obs. of  4 variables:
##  $ Leak    : chr  "No" "No" "No" "Yes" ...
##  $ Size    : num  1.75 1.75 1 1.5 1 1 1.75 1.75 1.5 1.75 ...
##  $ Material: chr  "Aldyl A" "Aldyl A" "Aldyl A" "Steel" ...
##  $ Pressure: chr  "High" "Med" "Low" "Med" ...
```

- OK, three `character` variables and one `numerical` variable.

R diversion: factor

- The first 10 elements of `Leak`

```
pipeline$Leak[1:10]
```

```
##  [1] "No"  "No"  "No"  "Yes" "No"  "Yes" "Yes" "No"  "No"  "No"
```

- Let's explicitly change `Leak` to a `factor` type and look at things again…

R diversion: factor

```
pipeline$Leak <- factor(pipeline$Leak)
str(pipeline)
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    1000 obs. of  4 variables:
##  $ Leak    : Factor w/ 2 levels "No","Yes": 1 1 1 2 1 2 2 1 1 1 ...
##  $ Size    : num  1.75 1.75 1 1.5 1 1 1.75 1.75 1.5 1.75 ...
##  $ Material: chr  "Aldyl A" "Aldyl A" "Aldyl A" "Steel" ...
##  $ Pressure: chr  "High" "Med" "Low" "Med" ...
```

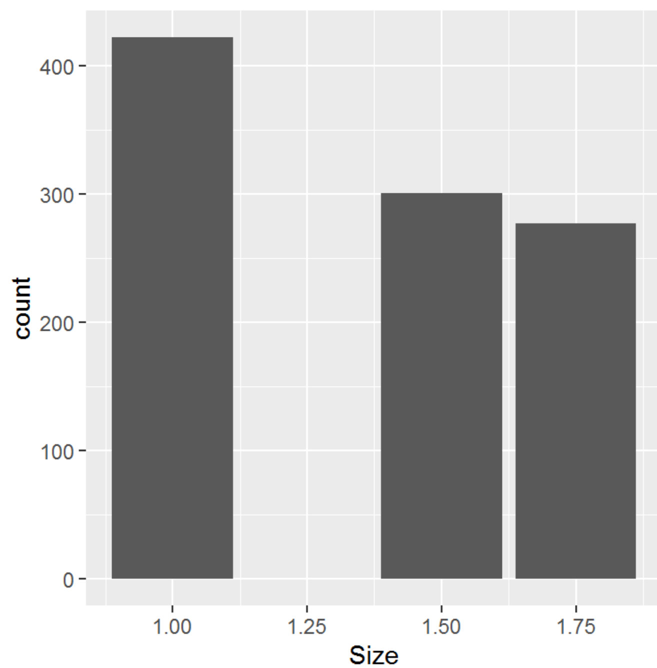
```
pipeline$Leak[1:10]
```

```
##  [1] No  No  No  Yes No  Yes Yes No  No  No 
## Levels: No Yes
```

- The tricky thing is that R will often, but not always, temporarily *coerce* variables to be `factor` type when it seems to make sense. This is in a sense related to the sometimes arbitrary division we make between *numerical* and *categorical*

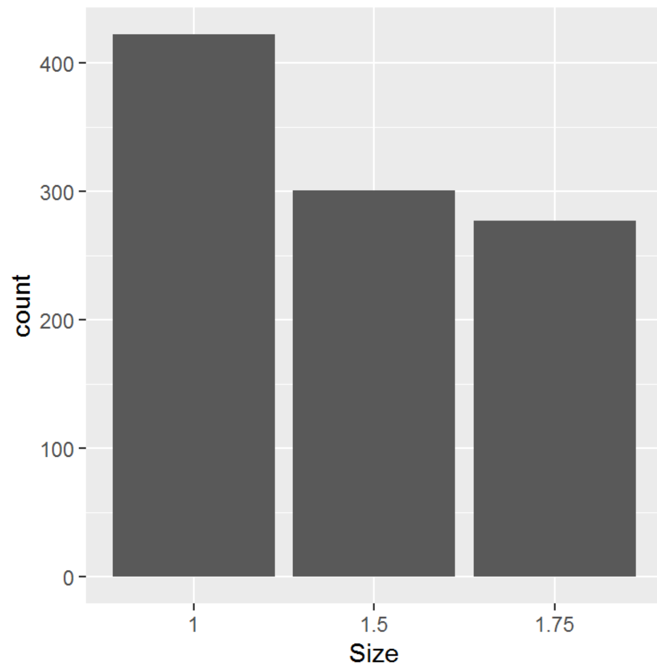
Plot where R guesses wrong

```
pipeline %>%
  ggplot(aes(x = Size)) + geom_bar()
```



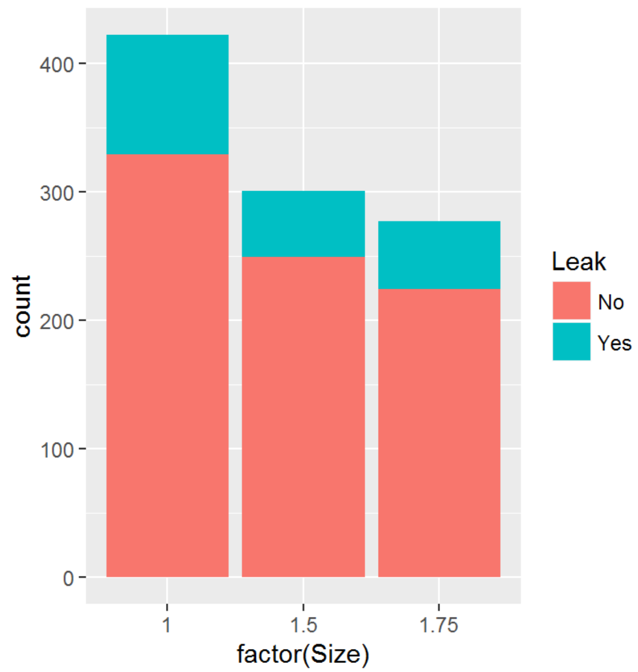
Explicitly declare a factor variable

```
pipeline %>%
  ggplot(aes(x = factor(Size))) + geom_bar() + xlab("Size")
```



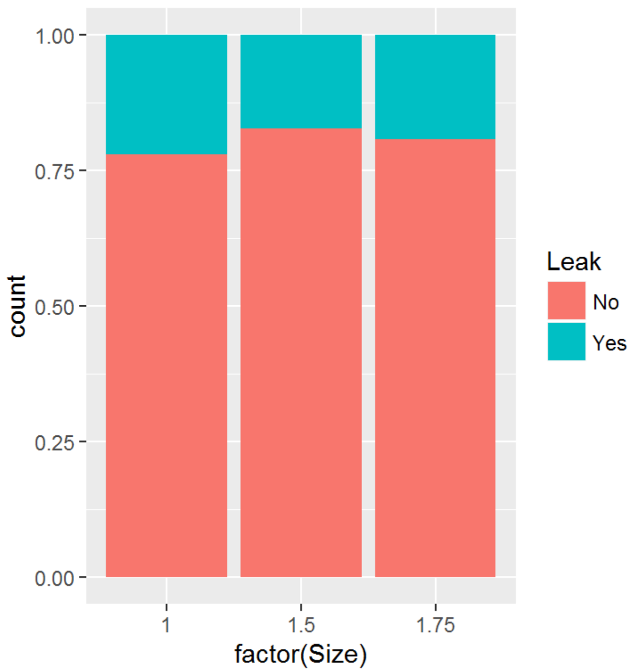
Plots for two-way classifications

- Stacked bar plot

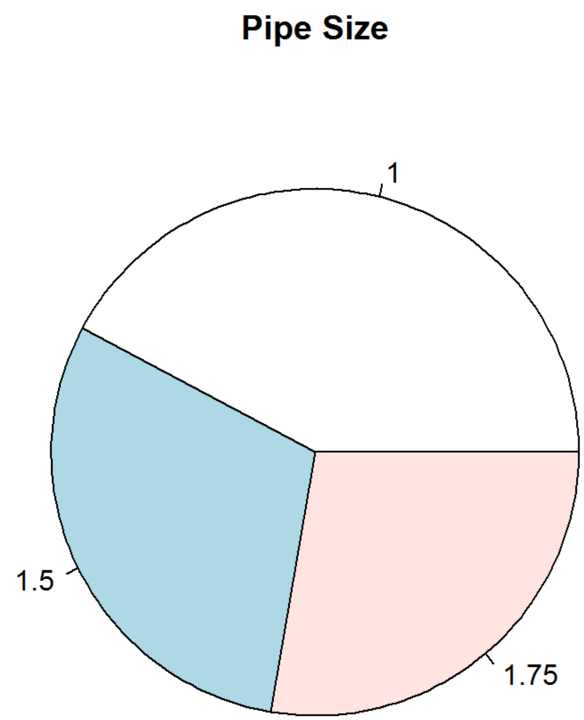


Plots for two-way classifications

- Not sure what this is called, but with proportions rather than counts



Crappy plot from hell that deserves to die



Inference for one-way classifications

Old model: population $X \sim N(\mu, \sigma^2)$

New model:

$$X = \begin{cases} 1 & \text{with probability } p, \\ 0 & \text{with probability } 1 - p \end{cases}$$

Or: $X \sim \text{Bernoulli}(p)$ (feel the Bern!)

Recall: $E(X) = p$ and $Var(X) = p(1 - p)$.

We don't know p. So as usual we (plan to) gather a sample $X_1, X_2, \ldots, X_n$

How to estimate p?

- Same as before: use $\overline{X}$, for suppose k is the number of "successes" (or "1"s) and think about how k/n is the same as $\overline{X}$.
- The traditional notation for this case is $\hat{p}$, but it's nothing more than $\overline{X}$.
- All the usual results follow:

$$E\left(\hat{p}\right) = p$$
$$Var(\hat{p}) = \frac{p(1-p)}{n}$$

- Getting close to the "Key Fact". We have the usual:

$$\frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \sim^{\text{approx.}} N(0, 1)$$

Confidence interval for a proportion

- As usual *Estimator +/- "2" times standard error*:

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}$$

  except for the usual problem…

- Simplest fix is to just replace p with its estimate $\hat{p}$.
- (Optional material) Actually a really badly performing confidence interval.

Another confidence interval for a proportion (OPTIONAL)

- The simplest interval is based on the approximation:

$$1 - \alpha = P\left(-z_{\alpha/2} < \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} < z_{\alpha/2}\right) \approx P\left(-z_{\alpha/2} < \frac{\hat{p} - p}{\sqrt{\frac{\hat{p}(1-p)}{n}}} < z_{\alpha/2}\right)$$

- A better performing interval is based on solving for p directly:

$$-z_{\alpha/2} < \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} < z_{\alpha/2}$$

Another confidence interval for a proportion (OPTIONAL)

Solution isn't so simple:

$$\frac{\hat{p} + \frac{z_{\alpha/2}^2}{2n}}{1 + \frac{z_{\alpha/2}^2}{n}} \pm \frac{z_{\alpha/2}}{1 + \frac{z_{\alpha/2}^2}{n}} \sqrt{\frac{\hat{p}\hat{q}}{n} + \frac{z_{\alpha/2}^2}{4n^2}}$$

But for the 95% interval our $z_{\alpha/2}$ is essentially "2", and the above pretty much reduces to *add two "successes" and two "failures"*:

$$\tilde{p} = \frac{k+2}{n+2}$$

The interval $\tilde{p} \pm z_{\alpha/2}\sqrt{\frac{\tilde{p}(1-\tilde{p})}{n}}$ performs better.

Example

- The 95% interval for the proportion of mains with leaks…

Model assumptions

- As usual depends on the skewness of the underlying distribution…
- Use the heuristic: $n\hat{p}$ and $n\hat{p}(1 - \hat{p})$ both exceed 5.

Inference for two-way classifications

```
##      Size
## Leak    1 1.5 1.75
##   No  329 249  224
##   Yes  93  52   53
```

- Main problem is to test if the rows and columns are either:
    - independent (language used when row *and* column totals are random)
    - "homogeneous" (language used when row *or* column totals are fixed in advance)
- Math is the same either way. So we'll focus on the question of "independence"
- Null hypothesis of the test: *rows and columns are independent*.

What does independence mean in this context?

Independence is a property of two random variables. In this case, two discrete random variables. The "model" is (for example in a 2x3 table)

Row/Column	1	2	3	Row margin
1	π_{11}	π_{12}	π_{13}	$\pi_{1\cdot}$
2	π_{21}	π_{22}	π_{23}	$\pi_{2\cdot}$
Column margin	$\pi_{\cdot 1}$	$\pi_{\cdot 2}$	$\pi_{\cdot 3}$	1

The null hypothesis is then technically:

$$\pi_{ij} = \pi_{i\cdot}\pi_{\cdot j} \text{ for all } i, j$$

Method

- Treat marginal totals as fixed
- Compute *expected cell counts* E_i assuming independence
- With O_i as *observed cell counts* use the following test statistic:

$$\sum_i \frac{(O_i - E_i)^2}{E_i} \sim^{\text{approx}} \chi^2_{(r-1)(c-1)}$$

- Approximation is good if $E_i \geq 5$ for all i.