Ch. 8: Inference for Categorical Data

MTH 361: Probability and Statistics in the Health Sciences

November 14, 2023

Announcements

Binomial data

Example: Advanced melanoma is an aggressive form of skin cancer that, until recently, was almost uniformly fatal. Two therapies have seemed to be successful in triggering an immune response to this cancer: nivolumab and ipilimunab.

A 2013 report in the New England Journal of Medicine by Wolchok et al. reported the results of a study in which patients were treated with both nivolumab and ipilimumab. Fifty-three patients were given the new regimens concurrently, and the response to therapy could be evaluated in 52 of the 53. Of the 52 evaluable patients, 21 (40%) experienced a response according to commonly accepted criteria. In previous studies, the proportion of patients responding to one of these agents was 30% or less. How might one compare the new data to past results?

Binomial data

Binomial variable: a variable with only two possible outcomes (also called *binary variables*)

- The possible outcomes are referred to as **successes** (the outcome we care about) and **failures** (the outcome we don't).

When we have a binomial variable, our goal is usually to estimate the underlying **probability of success**, or **population proportion**,

$$p$$

Sampling distribution of $\hat{p}$

The **sampling distribution of $\hat{p}$** has:

- Mean: np
- Standard error: $SE_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$
- Shape: Approximately normally distributed, as long as the sample size is "large enough"

$$np \geq 10$$

$$n(1-p) \geq 10$$

Sampling distribution of $\hat{p}$

We can calculate confidence intervals and carry out hypothesis tests using this approximation...

... but we like to be accurate.

So instead, we'll use methods based on the **binomial distribution**.

Binomial distribution

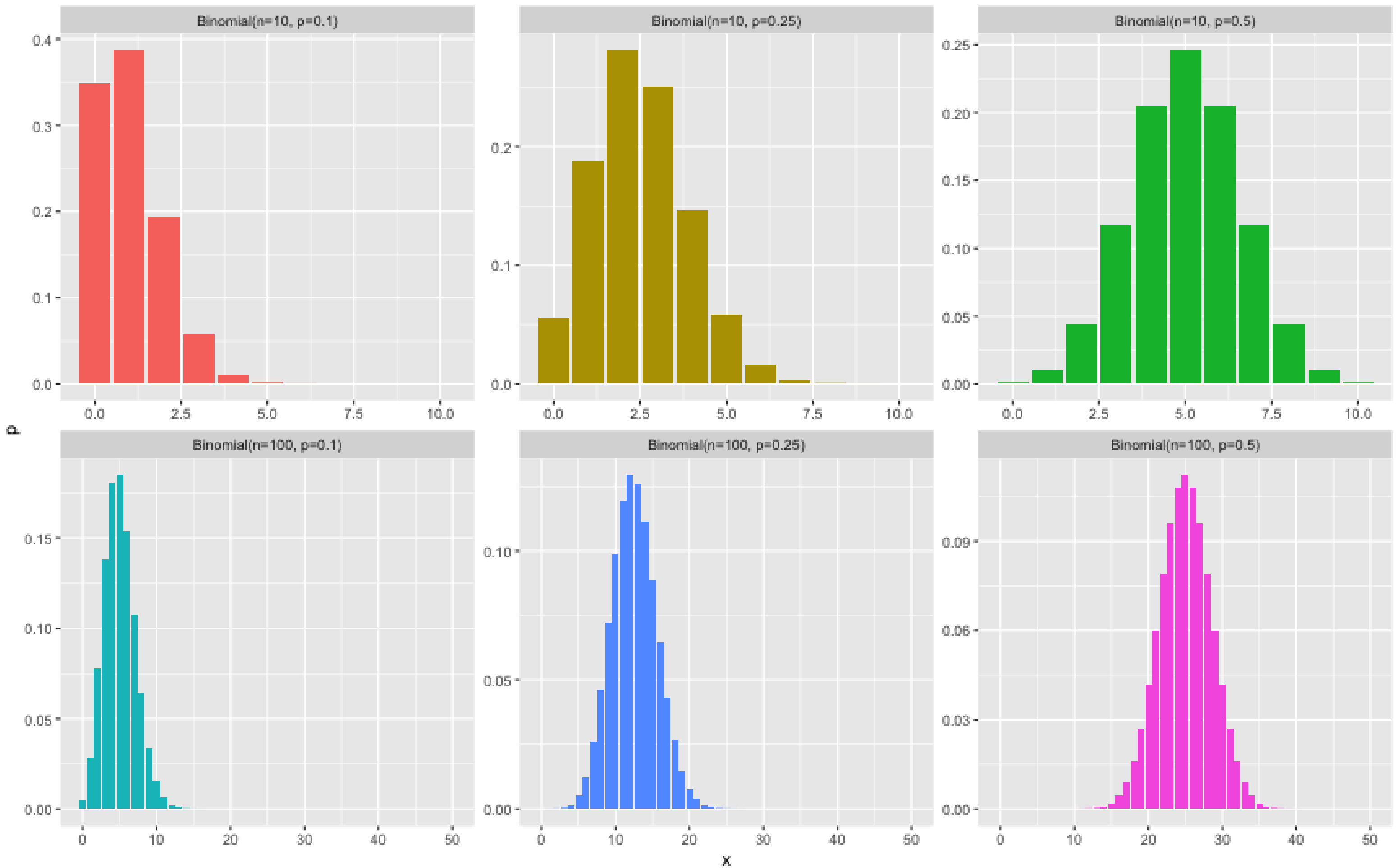
Assuming each observation is independent, we can get the formula for the **binomial distribution** from the Multiplication Rule. Let $P(X = x)$ represent the probability that we have observed x successes our of n trials.

$$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}$$

The binomial distribution is the "backbone" of the (1) binomial test, and (2) binomial interval.

Binomial distribution

What does the binomial distribution “look like”? It depends on the parameters.



Binomial test

Step 1: Write the null and alternative hypotheses

Assume that the true population proportion $p = p_0$, where p_0 is some constant value. The hypotheses we're interested in testing are:

$$H_0 : p = p_0$$

$$H_A : p(<, >, \neq)p_0$$

From our example:

$$H_0 : p = 0.30$$

$$H_A : p > 0.30$$

Binomial test

Step 2: Identify the test statistic

The data we'll need from our sample is the number of successes x and the sample size n.

Step 3: Calculate the p-value

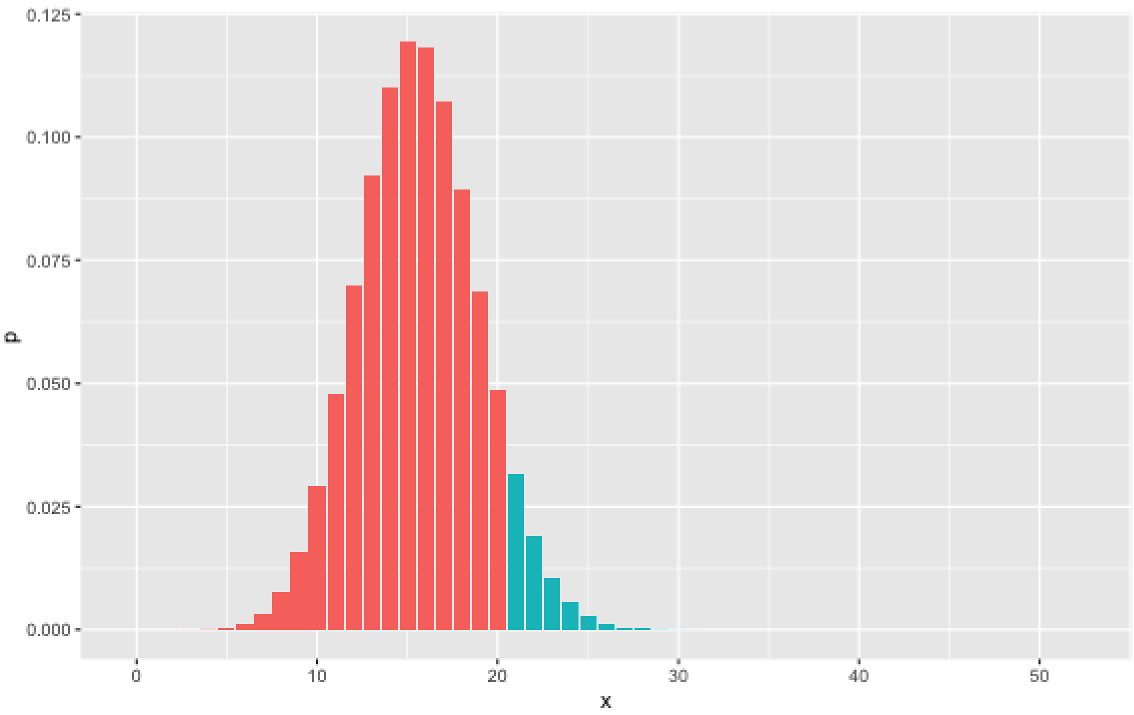
We’ll use the binomial distribution with $p = p_0$ to calculate the p-value.

1. Build a binomial distribution based on $H_0 : p = p_0$.

1. Calculate the probability of observing a "more extreme" result using this binomial distribution.

Binomial test

We have two "test statistics": the number of patients who improved (x=21), and the number of trials (n=52).



What's the probability that 21 patients or more would improve, **if the true improvement rate is 30%**?

Binomial test

Instead of calculating probabilities directly, we can use a function in R called `binom.test()`.

```
##
##
##
## data:  21 out of 52
## number of successes = 21, number of trials = 52, p-value = 0.07167
## alternative hypothesis: true probability of success is greater than 0.3
## 95 percent confidence interval:
##  0.2889045 1.0000000
## sample estimates:
## probability of success
##              0.4038462
```

Step 4: Make a conclusion to the research question

- There may be *strong* evidence to reject H_0 if $p - value < \alpha$.
- There may *not* be evidence to reject H_0 if $p - value > \alpha$.

Remember that sample size, data quality, choice of α, and consequences matter!

Binomial test

Example: Male radiologists have long suspected that they tend to have fewer sons than daughters. In a random sample of 87 children of "highly irradiated" male radiologists, 30 were male. Assume that the population proportion of male births is 0.519 (in the human population male babies are slightly more likely than female babies). Is there significant evidence to show male radiologists are less likely to have male babies?

Binomial test

Is there significant evidence to show male radiologists are less likely to have male babies?

```
##
##
##
## data:  30 out of 87
## number of successes = 30, number of trials = 87, p-value = 0.0007905
## alternative hypothesis: true probability of success is less than 0.519
## 95 percent confidence interval:
##  0.0000000 0.4374992
## sample estimates:
## probability of success
##              0.3448276
```

Confidence interval for a proportion

We’ll consider two choices:

1. Standard “Wald” confidence interval: based on the normal approximation to the sampling distribution

2. Clopper-Pearson confidence interval: based on the binomial distribution

Wald confidence interval

Wald interval: calculated based on the Central Limit Theorem and the normal distribution

$$\hat{p} \pm 1.96 \times SE(\hat{p})$$

where

$$SE_{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

- The normal distribution approximation can be inaccurate when n is small or p is "extreme"

Wald confidence interval

Example: Calculate and interpret a Wald confidence interval for the proportion of male babies born to male radiologists. Remember $\hat{p} = 30/87$.

Clopper-Pearson confidence interval

A another alternative is the **Clopper-Pearson confidence interval**, which is based on the binomial distribution.

The idea here is to take all of the values of

$$H_0 : p = p_0$$

for which we fail to reject the null hypothesis, and set those as the confidence interval!

- Also called the exact binomial confidence interval, because it's based on exact probabilities from a binomial distribution.

Clopper-Pearson confidence interval

The Clopper-Pearson interval is difficult to calculate by hand, but easy for R.

```
##
##
##
## data:  30 out of 87
## number of successes = 30, number of trials = 87, p-value = 0.001209
## alternative hypothesis: true probability of success is not equal to 0.519
## 95 percent confidence interval:
##  0.2461396 0.4544136
## sample estimates:
## probability of success
##              0.3448276
```

- There are other confidence intervals we could use. As $n \to \infty$, they tend to converge.

Comparing intervals

Example: A 2013 report in the New England Journal of Medicine by Wolchok et al. reported the results of a study in which patients were treated with both nivolumab and ipilimumab. Fifty-three patients were given the new regimens concurrently, and the response to therapy could be evaluated in 52 of the 53. Of the 52 evaluable patients, 21 (40%) experienced a response according to commonly accepted criteria. In previous studies, the proportion of patients responding to one of these agents was 30% or less.

Calculate and interpret a Wald confidence interval and a Clopper-Pearson confidence interval. How do they compare?

```
##
##
##
## data:  21 out of 52
## number of successes = 21, number of trials = 52, p-value = 0.1288
## alternative hypothesis: true probability of success is not equal to 0.3
## 95 percent confidence interval:
##  0.2700597 0.5489842
## sample estimates:
## probability of success
##              0.4038462
```

Sampling distribution of $p_1 - p_2$

The difference in two sample proportions, $\hat{p}_1 - \hat{p}_2$ tends to follow a normal model when:

- Each of the two samples are random samples from a population
- The two samples are independent of each other
- The sample sizes are "large enough": collectively, $n_1 p_1,\ n_2 p_2,\ n_1(1-p_1),\ n_2(1-p_2) \geq 10$

The standard error of the difference is

$$SE_{\hat{p}_1 - \hat{p}_2} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

Confidence intervals for difference

Calculate and interpret a 95% confidence interval for the difference in deaths due to breast cancer.

Group	Died from breast cancer	Did not die from breast cancer
Mammogram	500	44,425
Control	505	44,405

Association Between Categorical Variables

Take another look at the tables from the previous example:

Group	Died from breast cancer	Did not die from breast cancer
Mammogram	500	44,425
Control	505	44,405

This is a **two-way table**, which summarizes the relationship between two categorical variables.

A natural question for a two-way table is whether there is an association with these two categorical variables. In our context, is there evidence that the screening method is associated with the outcome?

Chi-square test for independence

Chi-square test for independence: procedure for determining whether or not two categorical variables are associated

- H_0: The "row" variable is independent of the "column" variable
- H_A: There is an association between the "row" variable and the "column" variable

Chi-square distribution

The χ^2 (chi-square) test statistic is:

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

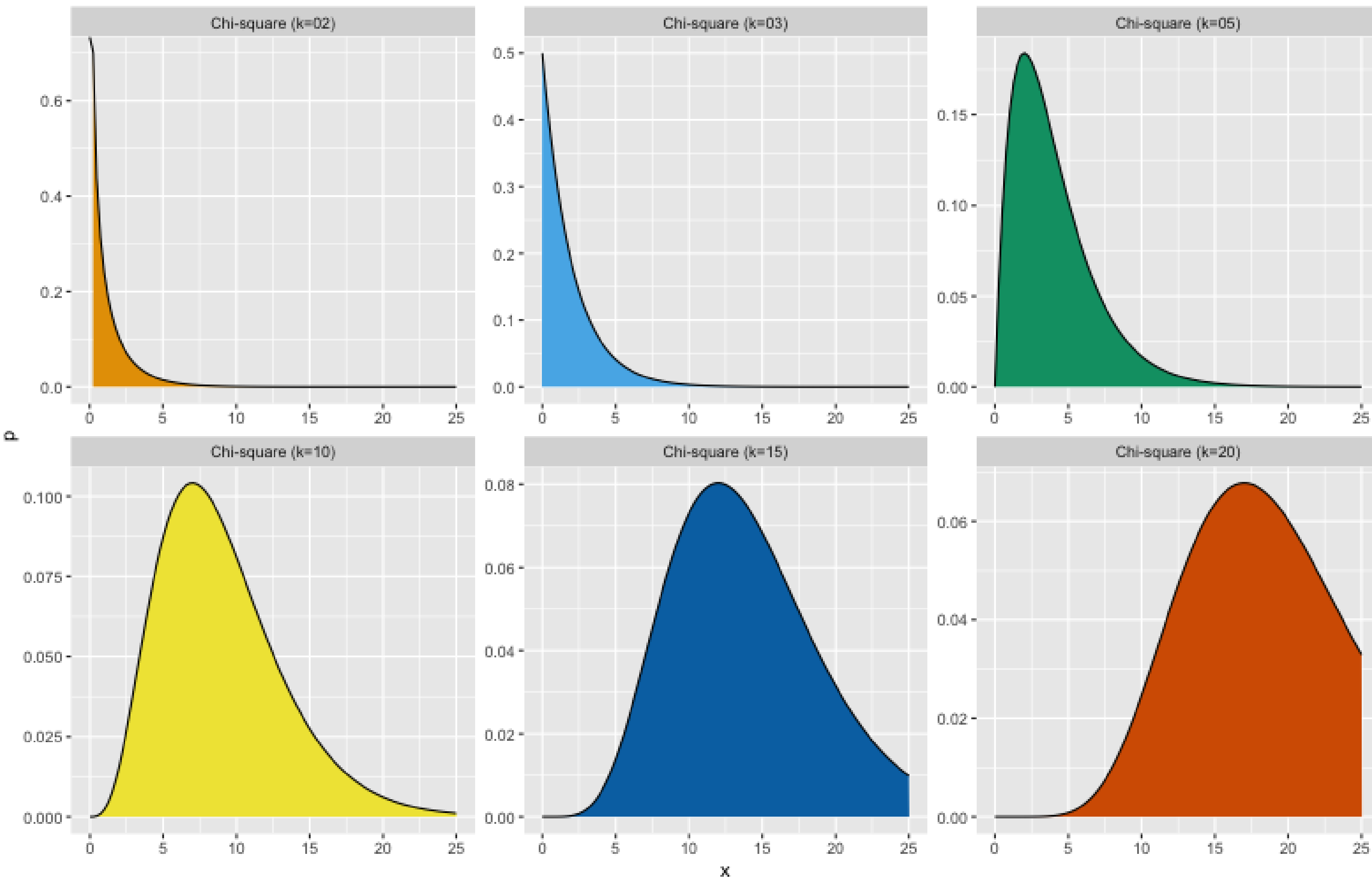
where k is the number of categories, O_i is the "observed" count in category i, and E_i is the "expected" count in category i under our model.

This test statistic follows a probability distribution called the χ^2 distribution:

- Defined on positive values only.
- Right-skewed probability distribution.
- The higher the number of categories, k, the higher the expected value of the test statistic.

Chi-square distribution

How does the chi-square distribution change as k changes?

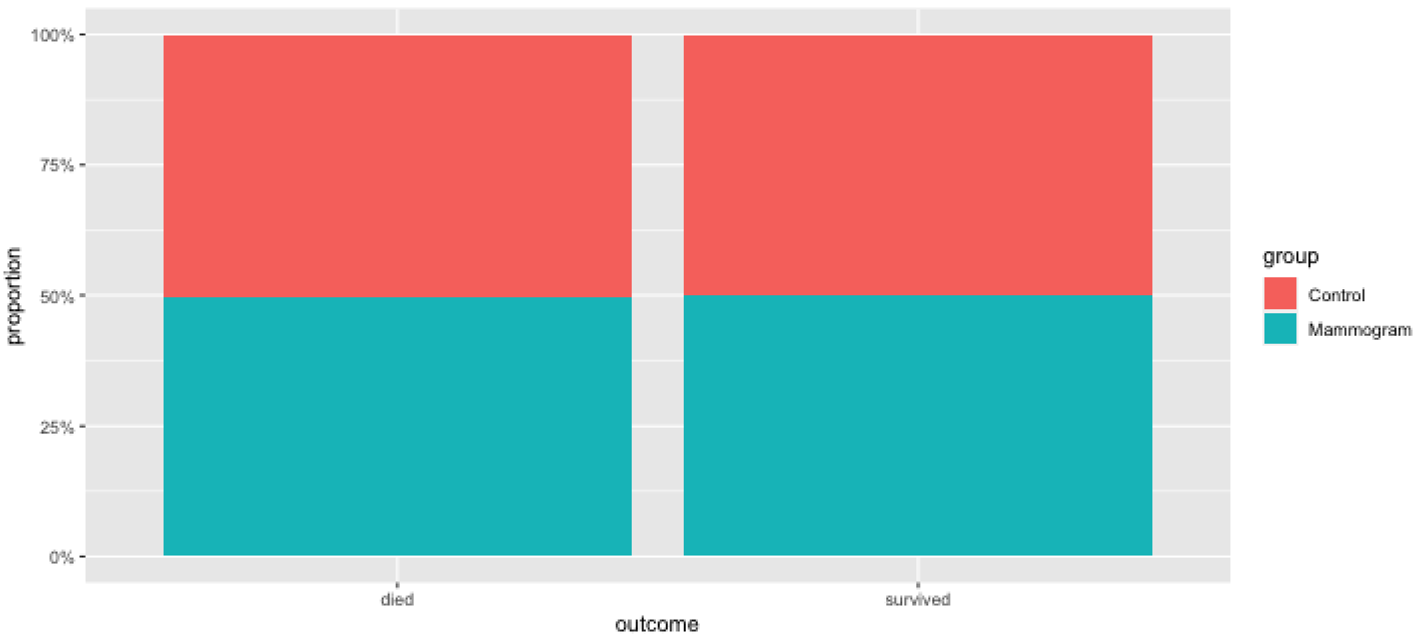


Back to the Example

Is there evidence that screening method is associated with outcome?

Group	Died from breast cancer	Did not die from breast cancer
Mammogram	500	44,425
Control	505	44,405

If there is no association, the proportions in each group should be approximately equal.



Chi-square test Calculation

First, let's expand the tables to include the total

Group	Died from breast cancer	Did not die from breast cancer	Survived
Mammogram	500	44,425	44,925
Control	505	44,405	44,910
Total	1005	88,830	89,835

The expected value E for a cell equals to

$$E = \text{row sum} * \text{column sum} \ / \ \text{total sum}$$

Some rule of thumb must be satisfied:

1. None of the categories should have an expected frequency < 1.

2. No more of than 20% of the categories should have expected frequency < 5.

These conditions help ensure the sample size is ”large enough”. Most statistical software, including R, will warn you if either rule of thumb may be violated.

Chi-square test

```
table <- tally(group~outcome, data=screening)
chisq.test(table)
```

```
##
##      Pearson's Chi-squared test with Yates' continuity correction
##
## data:  table
## X-squared = 0.01748, df = 1, p-value = 0.8948
```

Goodness-of-fit tests

There are a couple of ways to extend the 2*2 chi-square test

In general, *goodness-of-fit tests* are used to compare a data set to some known distribution.

Chi-square goodness-of-fit test: compares the observed frequencies in the data set to the expected frequencies under the null hypothesis or probability model

Test statistics and rule of thumb are exactly the same.

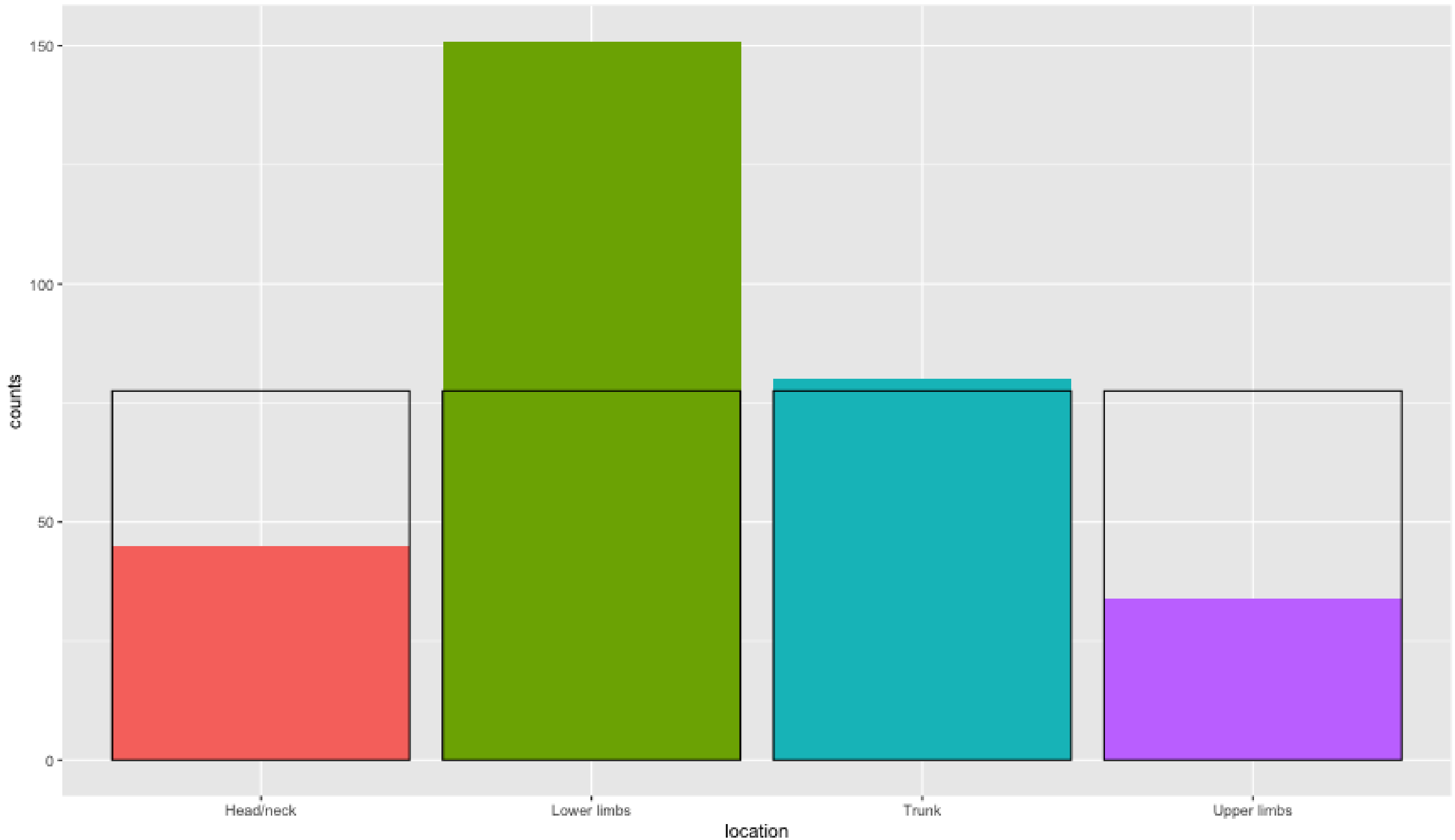
Chi-square goodness of fit

Example: Melanoma is a rare form of skin cancer that accounts for the great majority of skin cancer fatalities. UV exposure is a major risk factor for melanoma. Some body parts are regularly more exposed to the sun than others. A random sample of 310 women diagnosed with melanoma were classified according to the known location of the melanoma on their bodies.

Location	Head/neck	Trunk	Upper limbs	Lower limbs	Total
Count	45	80	34	151	310
Expected	77.5	77.5	77.5	77.5	

Assume that these body parts represent roughly equal skin areas. Do the data support the hypothesis that melanoma occurs evenly on the body?

Chi-square goodness of fit



Chi-square goodness of fit

Assume that these body parts represent roughly equal skin areas. Do the data support the hypothesis that melanoma occurs evenly on the body?

```
observed <- c(45, 80, 34, 151)
expected <- c(0.25, 0.25, 0.25, 0.25)
chisq.test(x=observed, p=expected)
```

```
##
##      Chi-squared test for given probabilities
##
## data:  observed
## X-squared = 107.83, df = 3, p-value < 2.2e-16
```