# Stat 5421 Notes: Review of Intro Stats with Some Extras

Charles J. Geyer

September 13, 2023

## Contents

## 1 License

## 2 What is This?

Chapter 1 of the textbook (Agresti) already reviews bits of the theory of probability and statistics, but IMHO that chapter is not basic enough for some students. Hence this.

This course is about analysis of discrete data, mostly count data. So most of intro stats is not relevant. This course also has as a prerequisite some course beyond intro stats. So we assume you are all more sophisticated than the typical student in a intro stats course. But just having had some course (a long list of options, no particular course required) doesn't mean you are familiar with any particular material not in intro stats.

So we start by reviewing the material in intro stats that is in what this course is about. Along the way we add a few alternative methods that intro stats omits, which will be prominent in this course.

# 3 Statistical Inference

It may be annoying that much of the content of this course can be summarized as *it's complicated.* In order to keep all of these complications organized, we will make a lot of lists dividing up these complications.

*Statistical inference* is the process of saying something about the probability distribution that describes the data. More on probability distributions later.

Statistical inference comes in two main kinds and several minor kinds. The major kinds are usually called

- frequentist and
- Bayesian.

The minor kinds will not be used in this course. The two kinds above may not have been mentioned in your previous statistics courses. If not, everything they did was frequentist, and they just ignored Bayesian. Hence you are expected to know something about frequentist, but may not know anything about Bayesian.

All of statistical inference has the problem that it can never get the right answer: estimators are equal to the parameters they estimate with very low probability. Hence every statistical inference comes with an indication of how wrong it may be. There are three main kinds of frequentist statistical inference

- point estimates,
- confidence intervals (sometimes called interval estimates), and
- tests of statistical hypotheses (hypothesis tests, for short, also called significance tests).

You have seen all of these in previous statistics courses.

There is only one kind of Bayesian inference. Bayesians say that probability theory is the correct way to describe uncertainty. So a probability distribution is the correct way to describe our uncertainty about the true unknown distribution of the data or any of its parameters. We are uncertain about the true unknown parameter value both before and after we see the data. But our uncertainties before and after are different. We know more, are less uncertain, after the data have been seen. The distribution that describes our uncertainty before is called the *prior* distribution (or just prior), the distribution that describes our uncertainty after is called the *posterior* distribution (or just posterior). The mathematical operation that connects prior, posterior, and data is called *Bayes' rule* or *Bayes' theorem.* It involves calculus and so will not be explicit in this course. You will learn all about that when you take a theory course. We will have a whole unit on Bayesian computation but very little theory. A taste of Bayesian inference is given in the section on Bayesian inference about a population proportion below.

# 4 Frequentist Inference

## 4.1 Point Estimates

Point estimates are often obvious, especially in intro stats.

- The sample mean is a natural estimate of the population mean.
- The sample median is a natural estimate of the population median.
- The sample proportion is a natural estimate of the population proportion.

And so forth. But often point estimates, especially in complicated models with multiple parameters,

- have no formula that expresses them   they can only be obtained by using numerical optimization with a computer   or
- any formula that expresses them is difficult to derive (uses calculus) and is not at all intuitive or obvious.

Examples of these start with these notes.

The most important thing to remember about point estimates is

- *point estimates are not the parameters they estimate* or
- $\hat{\theta}$ is not $\theta$.

The whole of statistical theory is about dealing with this issue. Every statistical estimate is wrong. And we have to quantify how wrong.

## 4.2 Hypothesis Tests

Hypothesis tests compare two statistical models (families of probability distributions). Formally they are called the null hypothesis and the alternative hypothesis. Tests come in three kinds:

- tests about the value of one parameter, which are further subdivided into
  - two-tailed tests,
  - lower-tailed tests, and
  - upper-tailed tests,
- tests comparing nested statistical models (which includes the bullet point above), and
- none of the above.

We will not deal with the last except to say that, in principle, a hypothesis test can compare any two hypotheses (logical statements) about the true unknown distribution of the data. But the only tests of practical interest are the first two kinds, and those are the only ones covered in statistics courses, including this one. So we will not mention "none of the above" again except when discussing model selection and model averaging, which is not, strictly speaking, hypothesis testing, but rather a competitor.

## 4.3 Intervals

Confidence intervals are only about one parameter. They do generalize to the notion of *confidence regions* for parameter vectors, but those are rarely used. Confidence intervals also come in two kinds:

- two-sided and
- one-sided,

but the latter are rarely used and not covered in most statistics courses. But we will cover them at the end of the course in the unit about inference when data are on the boundary.

## 4.4 Duality of Tests and Intervals

Hypothesis tests and confidence intervals are two ways of looking at the same math. Hypothesis tests can be used to make confidence intervals and vice versa.

- If you have a recipe for doing hypothesis tests at significance level $\alpha$ and for any null hypothesis $\theta$, then a confidence interval having coverage $1 - \alpha$ is the set of $\theta$ that the hypothesis test does not reject.
- Conversely, if you have a recipe for doing a confidence interval with coverage $1 - \alpha$, then a hypothesis test with null hypothesis $\theta$ and significance level $\alpha$ accepts the null hypothesis if $\theta$ is in the confidence interval and rejects it otherwise.

In more detail, two-tailed tests are dual to two-sided intervals. And two-tailed tests have null and alternative hypotheses

$$\begin{aligned} H_0 &: \text{true unknown parameter value} = \theta \\ H_1 &: \text{true unknown parameter value} \neq \theta \end{aligned}$$

So to make a confidence interval we (conceptually) perform the test for each $\theta$ in the parameter space, and the interval is the set of $\theta$ that the test accepts.

Thus, in theory, we have to do an infinite number of tests, but, in practice, we don't. We can just manipulate some formulas, or at worst solve a few equations numerically. Examples in the next section.

Going the other way is much simpler. The hypothesis test rejects the null hypothesis of equality to $\theta$ if and only if $\theta$ is not in the confidence interval.

Also in more detail, one-tailed tests are dual to one-sided intervals. A lower-tailed test has null and alternative hypotheses

$$\begin{aligned} H_0 &: \text{true unknown parameter value} = \theta \\ H_1 &: \text{true unknown parameter value} < \theta \end{aligned}$$

So to make a confidence interval we (conceptually) perform the test for each $\theta$ in the parameter space, and the interval is the set of $\theta$ that the test accepts.

Usually, if the test accepts a particular value of $\theta$, then it will also accept any value of $\theta$ greater than that. Thus a lower-tailed test is dual to a lower-bound confidence interval (the interval has the form $(l, \infty)$ for some lower bound $l$). Similarly, an upper-tailed test is dual to an upper-bound confidence interval (the interval has the form $(-\infty, u)$ for some upper bound $u$). If the parameter space $\Theta$ is not infinite in both directions, then

- a lower-bound interval has the form $\{\, \theta \in \Theta : l < \theta \,\}$ and
- an upper-bound interval has the form $\{\, \theta \in \Theta : \theta < u \,\}$.

If you are happy with what this section has said so far, you may skip to the next section, but if the description of null and alternative hypotheses for a one-tailed test is not what you saw in other courses, read on.

Some textbooks say the null and alternative hypotheses for a lower-tailed test are

$$\begin{aligned} H_0 &: \text{true unknown parameter value} \geq \theta \\ H_1 &: \text{true unknown parameter value} < \theta \end{aligned}$$

(we have changed $=$ to $\geq$ in the null hypothesis). Usually, these hypotheses give rise to exactly the same test procedure as the hypotheses given above. If they do not, then these hypotheses are a really dumb idea and one has no idea how to do the test.

The theoretical considerations centering around hypotheses of this form are discussed in these theoretical statistics lecture notes, pp. 181–185, but there is no need to look at that unless you are really curious about this issue.

## 4.5 Multiple Tests or Intervals

The mathematics of frequentist statistics is based on the following dogma.

- Do only one hypothesis test or confidence interval.
- Choose the procedure to be done before the data are collected.
- Do it, and report it.

Anything else is merely exploratory.

Doing multiple procedures and cherry picking the results you like is scientific fraud if you do not admit it and a red flag that your results are meaningless if you do admit it.

You can rigorously do multiple hypothesis tests or confidence intervals if you make adjustments to your procedures to account for that, which you may have seen in courses beyond intro stats and which we will cover in this course.

# 5 Inference About a Population Proportion

The binomial distribution may or may not have been explicitly described in the courses you have had previously. Many intro stats courses discuss this under the description *inference about a population proportion*. But all advanced courses, including this one, say *binomial distribution*. We will cover inference for the binomial distribution in exhaustive detail later. For now we just review what was said about it in intro stats, and we add a few alternative methods that were not covered in intro stats.

Let $\pi$ denote the population proportion of some quantity and $\hat{\pi}_n$ denote the sample proportion for sample size $n$. Then it is an instance of the central limit theorem that

$$\frac{\hat{\pi}_n - \pi}{\sqrt{\pi(1 - \pi)/n}} \tag{$*$}$$

has an approximate standard normal distribution (mean zero, variance one) when $n$ is sufficiently large. How large it needs to be depends on what the true unknown parameter value $\pi$ is, what question you are asking, and how accurate you want the approximation to be. The reason why this is an instance of the central limit theorem is that $\hat{\pi}_n$ is a sample mean.

Here $\pi$ is the unknown parameter, not $3.1415926\ldots$. In this we are following Agresti who follows the rule that parameters are always denoted by Greek letters (not mostly Greek letters, like some textbooks do).

In this course, there is no exact inference (except for fuzzy tests and confidence intervals) unlike what you have seen in courses about analyzing continuous data. For normally distributed data, you have t and F tests, which are exact (assuming exact homoscedastic normality). But there is nothing like that for categorical data. We have to make do with approximations, and this starts here.

*If you are trying to do a t or F test in this course, then you are making a mistake.*

## 5.1 Hypothesis Tests

### 5.1.1 The Usual Procedure

The generally accepted way to do a hypothesis test about a sample proportion is to use ($*$) as the test statistic with $\pi$ being the hypothesized value under the null hypothesis rather than the true unknown parameter value.

Suppose we want to test the null hypothesis $\pi = 1/2$ and the data are as follows.

```r
pi <- 1 / 2
n <- 10
x <- 8
pi.hat <- x / n
z <- (pi.hat - pi) / sqrt(pi * (1 - pi) / n)
z
```

```
## [1] 1.897367
```

```r
# upper-tailed p-value
pnorm(z, lower.tail = FALSE)
```

```
## [1] 0.02888979
```

```r
prop.test(x, n, p = pi, alternative = "greater", correct = FALSE)$p.value
```

```
## [1] 0.02888979
```

```r
# lower-tailed p-value
pnorm(z)
```

```
## [1] 0.9711102
```

```r
prop.test(x, n, p = pi, alternative = "less", correct = FALSE)$p.value
```

```
## [1] 0.9711102
```

```r
# two-tailed p-value
pnorm(- abs(z)) + pnorm(abs(z), lower.tail = FALSE)
```

```
## [1] 0.05777957
```

```r
2 * pnorm(abs(z), lower.tail = FALSE)
```

```
## [1] 0.05777957
```

```r
2 * pnorm(- abs(z))
```

```
## [1] 0.05777957
```

```r
2 * min(pnorm(z), pnorm(z, lower.tail = FALSE))
```

```
## [1] 0.05777957
```

```r
prop.test(x, n, p = pi, correct = FALSE)$p.value
```

```
## [1] 0.05777957
```

(The last four results involving `pnorm` are equal because of symmetry of the normal distribution. The `correct = FALSE` tells `prop.test` not to use its default correction for continuity, which few, if any, textbooks or experts recommend.)

The interpretation of the test is that, if we can believe an upper-tailed test is valid (if we can believe that this procedure was selected as the only hypothesis to be tested before the data were collected — or at least could have been  they weren't data snooping), then $P = 0.029$ is below the conventional 0.05 level for "statistical significance". But you might consider this weak, moderate, or strong evidence against the null hypothesis, depending on what the data are about, details of the experiment, and so forth.

If the authors of the test (or our second-guessing of them) consider the two-tailed test appropriate, then $P = 0.058$ is above the conventional 0.05 level for "statistical significance". But you might still consider this weak or moderate evidence against the null hypothesis, again depending on what the data are about, details of the experiment, and so forth.

Later on we will see that this test is a special case of a general procedure called the Rao test or the score test.

### 5.1.2 The Likelihood Ratio Test

We interrupt our review of intro stats to show you something you may not have seen before. This procedure is very widely used, and we will use it a lot in this course. You may have used it yourself. It is what R generic function `anova` does when handed the results of calls to R function `glm`. An application of large sample theory of statistics (the central limit theorem plus Taylor series approximation) says that

$$2 \left[ x \log \left( \frac{\hat{\pi}_n}{\pi} \right) + (n - x) \log \left( \frac{1 - \hat{\pi}_n}{1 - \pi} \right) \right] \tag{$**$}$$

has an approximate chi-square distribution distribution with one degree of freedom.

So another good way to do a hypothesis test about a sample proportion is to use (∗∗) as the test statistic with $\pi$ being the hypothesized value under the null hypothesis rather than the true unknown parameter value.

You might ask: who ordered that? Where does (∗∗) come from? The answer is from theory. More on this later.

Since this test statistic is large when $\hat{\pi}_n$ is far from $\pi$ in either direction (as shown by a plot below), this is inherently a two-tailed test. One-tailed tests are covered in the next section.

```
tstat <- 2 * (x * log(pi.hat / pi) + (n - x) * log((1 - pi.hat) / (1 - pi)))
# two-tailed hypothesis test
pchisq(tstat, df = 1, lower.tail = FALSE)
```

```
## [1] 0.04960103
```

These tests are asymptotically equivalent, meaning that for sufficiently large $n$ they give nearly the same results for the same data. Here they give somewhat different results, $P = 0.058$ for the usual (Rao) test and $P = 0.050$ for the likelihood ratio test.

**Warning:** the likelihood ratio test statistic calculation above blows up, giving `0 * -Inf = NaN` when either the number of successes $x$ or the number of failures $n - x$ is zero, but this is incorrect, just bad calculation. Since

$$x \log(x) \to 0, \qquad \text{as } x \to 0$$

the term in the sum having observed sucesses zero or observed failures zero should be zero, so to be careful we calculate

```
tstat <- 0
if (x > 0) tstat <- tstat + 2 * x * log(pi.hat / pi)
if (x < n) tstat <- tstat + 2 * (n - x) * log((1 - pi.hat) / (1 - pi))
# two-tailed hypothesis test
pchisq(tstat, df = 1, lower.tail = FALSE)
```

```
## [1] 0.04960103
```

Let's see what happens when we make the sample size larger

```
n <- 100
x <- 61
pi.hat <- x / n
z <- (pi.hat - pi) / sqrt(pi * (1 - pi) / n)
tstat <- 2 * (x * log(pi.hat / pi) + (n - x) * log((1 - pi.hat) / (1 - pi)))
# two-tailed p-value for usual (Rao) test
2 * pnorm(- abs(z))
```

```
## [1] 0.0278069
```

```
# two-tailed p-value for likelihood ratio (Wilks) test
pchisq(tstat, df = 1, lower.tail = FALSE)
```

```
## [1] 0.02717247
```

Closer. And the $P$-values get closer and closer as $n$ goes to infinity.

A crazy way to calculate the results above. We do this to show how these are special cases of the way R compares complicated models (general tests of model comparison done by R generic function `anova`)

```r
gout0 <- glm(cbind(x, n - x) ~ 0, family = binomial)
gout1 <- glm(cbind(x, n - x) ~ 1, family = binomial)
anova(gout0, gout1, test = "Rao")[2, "Pr(>Chi)"]
```

```
## [1] 0.0278069
```

```r
anova(gout0, gout1, test = "LRT")[2, "Pr(>Chi)"]
```

```
## [1] 0.02717247
```

These look odd, but that is just he way R function `glm` works. For binomial data,

- the response is either a zero-or-one-valued vector, or a two-column matrix of counts, first column successes, second column failures, so here we have a $1 \times 2$ matrix `cbind(x, n - x)`,
- in the formula 0 means no intercept and 1 means has intercept (which is the default, but we need the 1 here because there are no other terms in the formula),
- no intercept just happens to be the null hypothesis we want but we could test a different null hypothesis using the `offset` argument to R function `glm`.

### 5.1.3 The Signed Likelihood Ratio Test

Since the square of a standard normal random variable has the chi-squared distribution with one degree of freedom (by definition), the square root of a chi-square random variable with one degree of freedom has the same distribution as the absolute value of a standard normal random variable. If we attach a random sign, plus and minus being equally probable, to that square root, we get back a standard normal random variable. By $(*)$ being approximately standard normal, $\hat{\pi}_n - \pi$ is approximately equally likely to be plus or minus (when $\pi$ is the true unknown parameter value). Thus the square root of $(**)$ with the sign of $\hat{\pi}_n - \pi$ attached is approximately standard normal for large $n$.

Thus we have another test statistic

$$\operatorname{sign}(\hat{\pi}_n - \pi) \sqrt{2x \log\left(\frac{\hat{\pi}_n}{\pi}\right) + 2(n - x) \log\left(\frac{1 - \hat{\pi}_n}{1 - \pi}\right)} \tag{$***$}$$

which if squared gives back the test statistic $(**)$ of the likelihood ratio test. But now this test statistic has an approximately normal distribution and can be used for one-tailed tests.

Because the sign of the test statistic is the sign of $\hat{\pi}_n - \pi$, a lower-tailed test has a low $P$-value when $\hat{\pi}_n$ is a lot less than $\pi$, and an upper-tailed test has a low $P$-value when $\hat{\pi}_n$ is a lot greater than $\pi$, which is what these tests are supposed to do.

We return to our original data

```r
n <- 10
x <- 8
pi.hat <- x / n
tstat <- sign(pi.hat - pi) * sqrt(2 * (x * log(pi.hat / pi) +
    (n - x) * log((1 - pi.hat) / (1 - pi))))
# upper-tailed p-value
pnorm(tstat, lower.tail = FALSE)
```

```
## [1] 0.02480051
```

```r
# lower-tailed p-value
pnorm(tstat)
```

```
## [1] 0.9751995
```

```r
# two-tailed p-value
2 * pnorm(- abs(tstat))
```

```
## [1] 0.04960103
```

We included the two-tailed $P$-value only to show that a two-tailed signed likelihood ratio test gives exactly the same results as a likelihood ratio test.

### 5.1.4 The Wald Test

This section has no example code because this procedure is not recommended for hypothesis tests about a population proportion. We only have this section to mark that there is yet another general test procedure to learn.

The Wald test, if we did it (which we won't), would use the test statistic (†) in the following section but would otherwise calculate $P$-values just like the Rao test and the signed likelihood ratio test.

But that is not recommended by textbooks. The justification for using $(*)$ instead of (†) is that the $P$-value is supposed to be calculated using the assumption that $\pi$ is the true unknown parameter value, that is, the true unknown parameter value is the value hypothesized by the null hypothesis. Under that assumption the standard deviation of $\hat{\pi}_n - \pi$ is $\sqrt{\pi(1 - \pi)/n}$. Hence that should be used to standardize it, giving $(*)$ rather than (†).

Later on we will see that this test is a special case of a general procedure called the Wald test. When $n$ is large there is nothing wrong with Wald tests. If you have ever looked at the output of R function `summary` applied to the output of R function `lm` or `glm`, then you have looked at the $P$-values in the rightmost column. All come from Wald tests.

So if there were anything really wrong with Wald tests, there would be a big problem. But there isn't (when $n$ is large).

Here, of course, $n = 10$ is not "large".

## 5.2 Confidence Intervals

### 5.2.1 Usual Intervals

We now return to reviewing intro stats. Another application of large sample theory of statistics (the central limit theorem plus Slutsky's theorem, also called the plug-in principle), says that

$$\frac{\hat{\pi}_n - \pi}{\sqrt{\hat{\pi}_n (1 - \hat{\pi}_n)/n}} \tag{†}$$

has an approximate standard normal distribution when $n$ is sufficiently large (the only difference between $(*)$ and (†) is that the former has $\pi$ in the denominator where the latter has $\hat{\pi}_n$).

The concern that the Wald test does not use the null hypothesis in calculating standard deviation, which leads textbooks to (sometimes) avoid recommending it, does not apply to confidence intervals, because they have no null hypothesis.

We find the confidence interval dual to this test by inverting the test. A two-tailed test rejects the null hypothesis at level $\alpha$ when the absolute value of (†) is large. So we accept $\pi$ when

$$\left| \frac{\hat{\pi}_n - \pi}{\sqrt{\hat{\pi}_n (1 - \hat{\pi}_n)/n}} \right| < z_{\alpha/2}$$

where $z_{\alpha/2}$ denotes the upper $\alpha/2$ quantile of the standard normal distribution. The set of all accepted $\pi$ is the confidence interval with (approximate, large $n$) coverage $1 - \alpha$.

Solving

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

for $\pi$ gives

$$\hat{\pi}_n - z_{\alpha/2}\sqrt{\hat{\pi}_n(1 - \hat{\pi}_n)/n} < \pi < \hat{\pi}_n + z_{\alpha/2}\sqrt{\hat{\pi}_n(1 - \hat{\pi}_n)/n}$$

so we get a confidence interval with coverage $1 - \alpha$ having endpoints

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

Try that.

```
conf.level <- 0.95
crit <- qnorm((1 + conf.level) / 2)
crit
```

```
## [1] 1.959964
```

```
se.pi.hat <- sqrt(pi.hat * (1 - pi.hat) / n)
pi.hat + c(-1, 1) * crit * se.pi.hat
```

```
## [1] 0.552082 1.047918
```

(the `c(-1, 1)` is R for ±).

This interval is called a Wald interval, like the test it is dual to being called a Wald test.

Note that the upper endpoint of the confidence interval is outside the parameter space. But it means the same as if we had reported the confidence interval as (0.552, 1)

In addition to this minor blemish, this Wald interval for a population proportion has a very big problem. When $\hat{\pi}_n$ is equal to zero or one (when $x$ is equal to 0 or $n$) the interval has width zero, which is useless and stupid. But this happens with negligible probability when $n$ is large.

We continue this analysis with a section on fixing up Wald intervals below.

### 5.2.2 Score Intervals

Thus some textbooks recommend an interval that is harder to calculate. This inverts the test that uses ($*$) as the test statistic. This is harder. The interval is the set of $\pi$ that satisfy

$$\left| \frac{\hat{\pi}_n - \pi}{\sqrt{\pi(1 - \pi)/n}} \right| < z_{\alpha/2}$$

It is not obvious how to solve this, but square both sides to get rid of the absolute value giving

$$\frac{(\hat{\pi}_n - \pi)^2}{\pi(1 - \pi)/n} < z_{\alpha/2}^2$$

and multiply both sides by the left-hand side denominator giving

$$(\hat{\pi}_n - \pi)^2 < z_{\alpha/2}^2 \pi(1 - \pi)/n$$

Since both sides are quadratic in $\pi$, this gives us a quadratic equation to solve. We won't bother with the details (they are in these theory notes, slides 113—116). The endpoints of this interval are

$$\frac{\hat{\pi}_n + \frac{z_{\alpha/2}^2}{2n} \pm z_{\alpha/2}\sqrt{\frac{\hat{\pi}_n(1-\hat{\pi}_n)}{n} + \frac{z_{\alpha/2}^2}{4n^2}}}{\left(1 + \frac{z_{\alpha/2}^2}{n}\right)}$$

What a mess! But it does have some advantages over the Wald interval.

- It never has zero width.
- It is always contained in the parameter space.
- It is asymptotically equivalent to (no better, no worse, when $n$ is large) the Wald interval.

Try it.

```
prop.test(x, n, conf.level = conf.level, correct = FALSE)$conf.int
```

```
## [1] 0.4901625 0.9433178
## attr(,"conf.level")
## [1] 0.95
```

```
(pi.hat + crit^2 / (2 * n) + c(-1, 1) * crit *
    sqrt(pi.hat * (1 - pi.hat) / n + crit^2 / (4 * n^2))) / (1 + crit^2 / n)
```

```
## [1] 0.4901625 0.9433178
```

We used the complicated formula just to show that R function `prop.test` with argument `correct = FALSE` does indeed calculate this interval.

Like the hypothesis test this is dual to, this interval is called the score (or Rao) interval. It is sometimes also called the Wilson interval after the first person to propose this special case of the general concept.

Despite the complexity of this interval, some intro stats textbooks recommend it because of the bad performance (in some cases when $n$ is small) of the Wald interval.

### 5.2.3 Likelihood Intervals

No intro stats books that I know of recommend these intervals, but they are widely used and we will use them in this course. They are obtained by inverting the test having test statistic ($**$). The interval is the set of $\pi$ that satisfy

$$2 \left| x \log \left( \frac{\hat{\pi}_n}{\pi} \right) + (n - x) \log \left( \frac{1 - \hat{\pi}_n}{1 - \pi} \right) \right| < z_{\alpha/2}^2$$

(we could have written $\chi^2_\alpha$ instead of $z^2_{\alpha/2}$ where the former means the upper $\alpha$ quantile of the chi-squared distribution with one degree of freedom, because squaring a standard normal random variable to get a chi-squared random variable maps two tails onto one tail). But there is no way to solve this in closed form. We have to do it by numerical solution of equations (more on this later).

```
invlogit <- function(theta) 1 / (1 + exp(- theta))
library(MASS)
gout <- glm(cbind(x, n - x) ~ 1, family = binomial)
suppressMessages(confint(gout, level = conf.level)) |> invlogit()
```

```
##     2.5 %    97.5 %
## 0.5005809 0.9636378
```

This code is rather bizarre. The R function that does this, method `glm` of R generic function `confint` in R package `MASS` only does likelihood intervals for regression coefficients of generalized linear model fits so we have to fake it out by

- treating this problem as a logistic regression with no covariates (only an intercept),
- letting it make a likelihood interval for that intercept parameter,
- mapping this interval from the logit scale to the probability scale (more on logit in the next section).

Later we will see how to do this by a more direct but still complicated method (here and here and here and here).

What we said about score intervals, also applies to likelihood intervals.

- It never has zero width.
- It is always contained in the parameter space.
- It is asymptotically equivalent to (no better, no worse, when $n$ is large) the Wald interval or the score interval.

So all three intervals are equally good for sufficiently large $n$ and the score and likelihood intervals are equally good for all $n$ (although they only have close to their nominal coverage for large $n$).

Since the computing here was completely mysterious, we show that it worked by plotting (∗∗) as a function of $\pi$ and showing that we have correctly computed the interval.

```r
foo <- function(pi) 2 * (x * log(pi.hat / pi) +
    (n - x) * log((1 - pi.hat) / (1 - pi)))
crit <- qchisq(conf.level, df = 1)
curve(foo, from = 0.4, to = 1, n = 1001, ylim = c(0, 1.5 * crit),
    ylab = expression(LRT(pi)), xlab = expression(pi))
abline(h = crit, lty = "dashed")
foo <- suppressMessages(confint(gout, level = conf.level)) |> invlogit()
abline(v = foo, lty = "dashed")
```

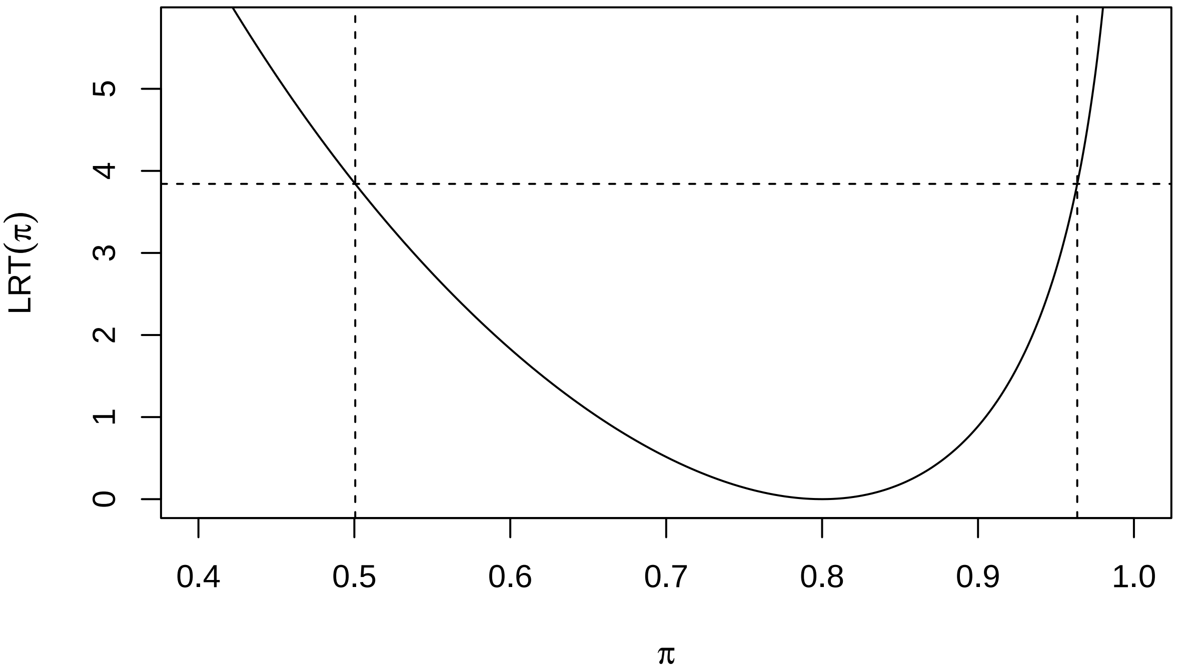

Figure 1: Likelihood Ratio Test Statistic Function and Likelihood Interval. Solid curve is (∗∗) plotted as function of $\pi$. Horizontal dashed line is at chi-squared critical value. Vertical dashed lines are at the endpoints of the calculated likelihood interval.

### 5.2.4 Fixing Up the Wald Interval

#### 5.2.4.1 Fixing Up Interval Goes Outside Parameter Space

The trick used in the preceding section, where we made a confidence interval for one parameter and then mapped that to a confidence interval for another parameter can be used to make sure Wald intervals are contained in the parameter space.

The new parameter

$$\theta = \operatorname{logit}(\pi) = \log\left(\frac{\pi}{1 - \pi}\right)$$

is called the *canonical parameter* of the binomial distribution. This is a term from the theory of exponential families of distributions. This function is so important that it has a special name, pronounced with a soft g: low-jit. It is also the default link function for `family = binomial` for R function `glm`, which is why it was used in the preceding example.

This function is invertible, and its inverse function is, as was seen in the preceding example,

$$\pi = \operatorname{logit}^{-1}(\theta) = \frac{\exp(\theta)}{1 + \exp(\theta)} = \frac{1}{1 + \exp(-\theta)}$$

As $\pi$ goes from zero to one, $\theta$ goes from $-\infty$ to $+\infty$. So the interval for $\theta$ can never be outside the parameter space for $\theta$, and must map into the parameter space for $\pi$.

So we make a confidence interval for $\theta$, which R function `glm` knows how to do, and then transform that to being a confidence interval for $\pi$.

```
crit <- qnorm((1 + conf.level) / 2)
pout <- predict(gout, se.fit = TRUE)
(pout$fit + c(-1, 1) * crit * pout$se.fit) |> invlogit()
```

```
## [1] 0.4592938 0.9495868
```

This trick of putting the endpoints of an interval through a function to give another interval only works when the function is monotone. But every scalar-to-scalar invertible function is monotone. So we do not need to check that.

#### 5.2.4.2 Fixing Zero-Width Wald Intervals

As mentioned above, Wald intervals have zero width when the data are on the boundary of the sample space. If $x = 0$, then $\widehat{\pi}_n = 0$. If $x = n$, then $\widehat{\pi}_n = 1$. In either case, $\widehat{\pi}_n(1 - \widehat{\pi}_n) = 0$ so the Wald interval has zero width.

Geyer (2009, DOI: 10.1214/08-EJS349) proposed a fixup for this. It works for all of the statistical models discussed in this course. But for binomial data it simplifies to the following.

- If $x = 0$, then a $1 - \alpha$ confidence interval is $0 < \pi < 1 - \alpha^{1/n}$.
- If $x = n$, then a $1 - \alpha$ confidence interval is $\alpha^{1/n} < \pi < 1$.

## 5.3 Summary of Frequentist Inference for Proportions

There is no one right way to do frequentist statistics. Just for two-tailed tests and two-sided intervals we had

- the usual (or Rao or score) test, $P = 0.058$,
- the likelihood ratio (or Wilks) test, $P = 0.050$,
- the Wald test, $P = 0.018$ (calculations not shown),
- the usual (or Wald) confidence interval, $(0.552, 1.048)$,

- the Wald confidence interval for $\theta$ mapped to an interval for $\pi$, (0.459, 0.950),
- the score (or Rao) confidence interval, (0.490, 0.943), and
- the likelihood (or Wilks) confidence interval, (0.501, 0.964).

But the dogma of only one procedure says you only do one of these, and you choose which one to do before seeing the data (unless there is a correction for doing more than one).

## 5.4 Bayesian Inference

Thomas Bayes (Wikipedia article) died in 1761 by which time he had written an unpublished note about the binomial distribution and what would now be called Bayesian inference for it using a flat prior. The note was found by a friend and read to the Royal Society of London in 1763 and published in its *Philosophical Transactions* in 1764 thus becoming widely known.

### 5.4.1 Prior

We have learned a little about Bayesian inference since then, and we now would rather allow any prior distribution that is a beta distribution (Wikipedia article).

This is mathematically convenient because it then follows that the posterior distribution is also a beta distribution. A family that has this property (if the prior is in the family, then so is the posterior) is called the *conjugate family of distributions* to the distribution of the data. So here the beta family of distributions (for the parameter) is conjugate to the binomial family of distributions (for the data).

The beta distribution is a distribution on the interval $(0, 1)$ that has two strictly positive parameters $\alpha_1$ and $\alpha_2$ that control the mean and variance of the distribution, but not in any obvious way. Moreover, if $f_{\alpha_1,\alpha_2}$ denotes the probability density function, then

$$\lim_{x \to 0} f_{\alpha_1,\alpha_2}(x) = \begin{cases} \infty, & 0 < \alpha_1 < 1 \\ \alpha_2, & \alpha_1 = 1 \\ 0, & 1 < \alpha_1 < \infty \end{cases}$$

and, similarly,

$$\lim_{x \to 1} f_{\alpha_1,\alpha_2}(x) = \begin{cases} \infty, & 0 < \alpha_2 < 1 \\ \alpha_1, & \alpha_2 = 1 \\ 0, & 1 < \alpha_2 < \infty \end{cases}$$

so the behavior is qualitatively different near the edges of the sample space for different values of $\alpha_1$ and $\alpha_2$. Nevertheless, $f_{\alpha_1,\alpha_2}(x)$ is continuous in $\alpha_1$, $\alpha_2$, and $x$ for $0 < \alpha_1 < \infty$ and $0 < \alpha_2 < \infty$ and $0 < x < 1$. It is only the limits above that are discontinuous.

For $\alpha_1 = \alpha_2 = 1$, the beta distribution is uniform (flat) on the interval $(0, 1)$. Thus our allowing any beta distribution includes the original calculation made by Bayes.

Recall that in Bayesian inference, it is the parameters that are considered random (and the data not random after they are seen). Thus these beta distributions are (in this Bayesian application) considered distributions of the parameter $\pi$. That is why we want distributions whose sample space is the parameter space for the binomial distribution (the interval $0 < \pi < 1$).

It is confusing when the prior and posterior distribution, which are distributions of the parameters of the given problem (here $\pi$), themselves have parameters (here $\alpha_1$, and $\alpha_2$). To avoid confusion, the latter are called *hyperparameters*. So

- $\pi$ is the parameter and
- $\alpha_1$ and $\alpha_2$ are the hyperparameters.

Bayesians treat parameters and hyperparameters very differently.

- Parameters are considered random. Their distributions (prior and posterior) describe our uncertainty about their true values.
- Hyperparameters are not considered random. They just say which prior and posterior we are talking about.

### 5.4.2 Posterior

When applied to this problem Bayes' rule says,

- if the prior is the beta distribution with hyperparameters $\alpha_1$ and $\alpha_2$,
- then the posterior is the beta distribution with hyperparameters $x + \alpha_1$ and $n - x + \alpha_2$,

where (as above) the binomial data are $x$ successes in $n$ trials.

It takes a large amount of theory to derive this, which is beyond the scope of this course. It is done in theory courses (for example these notes, slides 6—12).

Now our Bayesian inference is a whole distribution (the posterior). We can show it by plotting its probability density function. We start with the $\alpha_1 = \alpha_2 = 1$ case that Bayes used

```r
alpha1 <- 1
alpha2 <- 1
curve(dbeta(pi, x + alpha1, n - x + alpha2), from = 0, to = 1,
    xname = "pi", xlab = expression(pi), ylab = "probability density")
curve(dbeta(pi, alpha1, alpha2), from = 0, to = 1,
    xname = "pi", add = TRUE, lty = "dashed")
```

### 5.4.3 Concentration

Hopefully, you can just see from the picture that the posterior is more concentrated than the prior (the prior is more spread out).

If not, the variance of the posterior distribution is

$$\frac{(x + \alpha_1)(n - x + \alpha_2)}{(n + \alpha_1 + \alpha_2)^2 (n + 1 + \alpha_1 + \alpha_2)} \approx \frac{\hat{\pi}_n (1 - \hat{\pi}_n)}{n}$$

for large $n$. Thus we see that when $n$ is large

- the prior does not matter and
- the Bayesian (approximate) posterior variance is the same as the frequentist (approximate) sampling variance.

These two facts are a special case of a general phenomenon called the Bernstein—von Mises theorem (Wikipedia page).

We also see that the Bayesian agrees with the frequentist about the square root law: statistical precision varies as the square root of the sample size.

### 5.4.4 Bayesian Learning Paradigm

Prior and posterior distributions are the same sort of thing. In fact a prior can be a posterior.

A distribution can be the posterior distribution incorporating some data we have seen but also the prior distribution for analyzing some data we have not yet seen.

The posterior for the previous analysis serves as the prior for our next analysis. Whatever you call it, it describes our current uncertainty about the true unknown parameter value. Suppose we observe some more data.

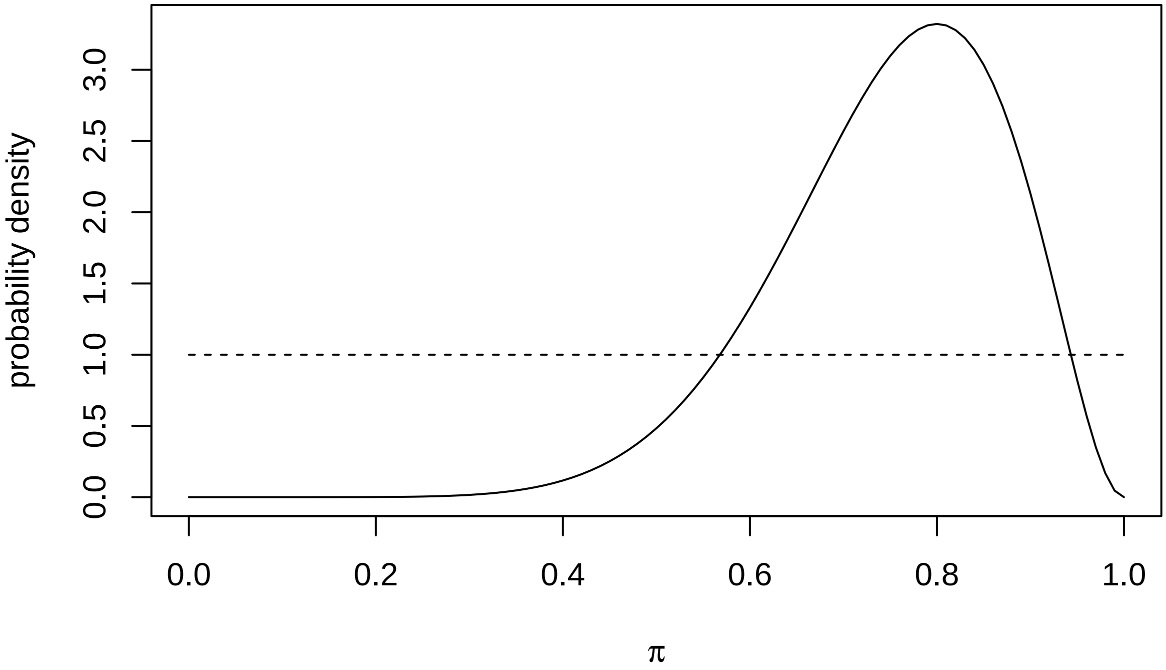

Figure 2: Binomial data with x = 8 and n = 10. Solid curve posterior, dashed curve prior.

```
# update hyperparameters incorporating old data
alpha1 <- x + alpha1
alpha2 <- n - x + alpha2
# new data
x <- 3
n <- 7
```

```
curve(dbeta(pi, x + alpha1, n - x + alpha2), from = 0, to = 1,
    xname = "pi", xlab = expression(pi), ylab = "probability density")
curve(dbeta(pi, alpha1, alpha2), from = 0, to = 1,
    xname = "pi", add = TRUE, lty = "dashed")
```

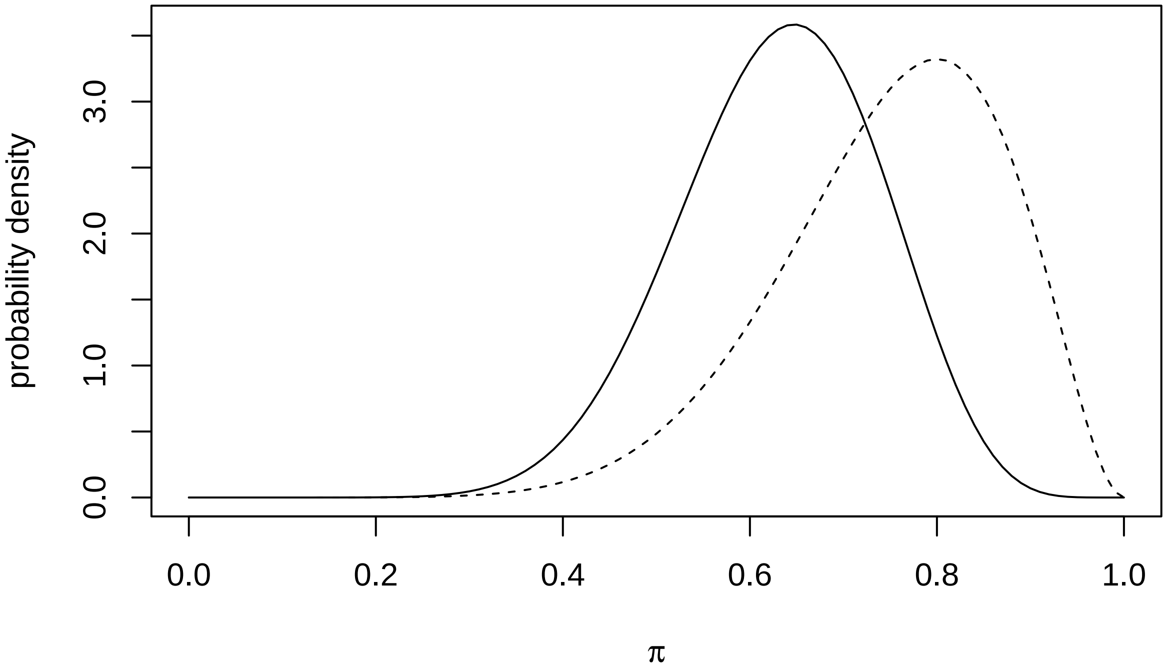

Figure 3: Posterior Distribution. Binomial data with x = 3 and n = 7. Solid curve posterior, dashed curve prior.

Note that we can write the hyperparameters more symmetrically as

- $x + \alpha_1 =$ number of successes $+ \alpha_1$ and
- $n - x + \alpha_2 =$ number of failures $+ \alpha_2$.

Suppose we observe a sequence of data obtained in batches with $x_i$ successes and $y_i$ failures in batch $i$. Then

- for the first batch.
  - prior has hyperparameters $\alpha_1$ and $\alpha_2$ and
  - posterior has hyperparameters $x_1 + \alpha_1$ and $y_1 + \alpha_2$
- for the second batch.
  - prior has hyperparameters $x_1 + \alpha_1$ and $y_1 + \alpha_2$ (same as previous posterior)

    - posterior has hyperparameters $x_2 + x_1 + \alpha_1$ and $y_2 + y_1 + \alpha_2$
- for the third batch.
    - prior has hyperparameters $x_2 + x_1 + \alpha_1$ and $y_2 + y_1 + \alpha_2$ (same as previous posterior)
    - posterior has hyperparameters $x_3 + x_2 + x_1 + \alpha_1$ and $y_3 + y_2 + y_1 + \alpha_2$
- and so forth.

At each stage the posterior for old data is the prior for new data. And *this is very important* we get the same results for the same data no matter how we subdivide the data into batches.

All Bayesian inference works this way, not just this particular example. If Bayesian inference did not have this property, it would not be so philosophically satisfying. Since it does, it can serve as a model for all learning. Data arrives. Uncertainty is reduced. And Bayes' rule says exactly how.

### 5.4.5 Point Estimates

Bayesians have posterior distributions, but sometimes one does not want a whole picture. Also, when there is more than one parameter, pictures get harder to draw.

So one can just compute some numbers, for example

```r
# current posterior hyperparameters
alpha1 <- x + alpha1
alpha2 <- n - x + alpha2
# posterior mean
alpha1 / (alpha1 + alpha2)
```

```
## [1] 0.6315789
```

```r
# posterior median
qbeta(0.5, alpha1, alpha2)
```

```
## [1] 0.6362862
```

```r
# posterior mode
((alpha1 - 1) / (alpha1 + alpha2 - 2)) |> max(0) |> min(1)
```

```
## [1] 0.6470588
```

```r
# posterior standard deviation
sqrt(alpha1 * alpha2 / (alpha1 + alpha2)^2 / (alpha1 + alpha2 + 1))
```

```
## [1] 0.1078626
```

```r
# posterior quartiles
qbeta(c(1, 3) / 4, alpha1, alpha2)
```

```
## [1] 0.559113 0.708996
```

```r
# posterior interquartile range
qbeta(c(1, 3) / 4, alpha1, alpha2) |> diff()
```

```
## [1] 0.149883
```

And this could keep on going forever. Hence the modern emphasis on posterior distributions rather than point estimates.

# 6 Inference About a Two-Dimensional Contingency Table

## 6.1 Contingency Tables

A *contingency table* is just a table of counts. They come in any dimension, 1, 2, 3, 4, .... One-dimensional is covered in these notes. Three-and-higher-dimensional is covered in these notes and in these notes.

## 6.2 Data

Here are some data http://www.stat.umn.edu/geyer/5421/mydata/multi-simple-3.txt.

We read them into R as follows. The data set read in by the R function `read.table` below (if this does not work on your computer see this announcement on the course home page).

```
foo <- read.table(
    url("http://www.stat.umn.edu/geyer/5421/mydata/multi-simple-3.txt"),
    header = TRUE)
class(foo)
```

```
## [1] "data.frame"
```

```
foo
```

```
##     y color    opinion
## 1  37   red       like
## 2  21  blue       like
## 3  25 green       like
## 4  37   red      so so
## 5  23  blue      so so
## 6  59 green      so so
## 7  49   red    dislike
## 8  29  blue    dislike
## 9  95 green    dislike
## 10 38   red no opinion
## 11 30  blue no opinion
## 12 57 green no opinion
```

These are data in the form that R function `glm` likes, a data frame. If you have heard of the tidyverse, it is based on the fundamental principle that all data should be put in this form (otherwise it is not "tidy").

But humans and some other R functions and the textbook (Agresti) like contingency tables. We turn it into that as follows.

```
bar  <- xtabs(y ~ color + opinion, data = foo)
class(bar)
```

```
## [1] "xtabs" "table"
```

```
bar
```

```
##        opinion
## color   dislike like no opinion so so
##   blue       29   21         30    23
##   green      95   25         57    59
##   red        49   37         38    37
```

In data frame form, we have three variables

- the count variable `y`,
- the categorical variable `color`, and

- the categorical variable `opinion`.

(in a data frame, every column is a variable).

In the table form

- the data are now in a two-dimensional array, and the count variable has no name, but corresponds to `y` in the data frame.
- the categorical variables in the data frame are also gone, they correspond to the row and column labels in the table.

```
dimnames(bar)
```

```
## $color
## [1] "blue"  "green" "red"
##
## $opinion
## [1] "dislike"    "like"       "no opinion" "so so"
```

- variable names in the data frame correspond to dimension names in the table and
- variable values in the data frame correspond to row and column labels in the table.

Since we have a contingency table with $3 \times 4 = 12$ cells, we also have that many possible parameters.

Since no one parameter is inherently more interesting than any other, we will not be interested in confidence intervals, only hypothesis tests.

Since there are a lot of parameters, it is not clear what hypotheses to test. The theory is complicated. The actual procedures are a lot simpler, so we will do practice before theory. This means we will not know how to interpret the tests until we have done the theory. You cannot interpret a hypothesis test until you are very clear about what the null hypothesis is.

## 6.3 R Function `chisq.test`

```
baz <- chisq.test(bar)
baz
```

```
##
##  Pearson's Chi-squared test
##
## data:  bar
## X-squared = 15.371, df = 6, p-value = 0.01756
```

Whatever the null hypothesis is (we don't know that yet), we can reject it at level = 0.05.

R function `chisq.test` likes the table form of the data.

To go further, we need the table of estimated cell mean values, also called expected values, or just *expected* for short.

```
moo <- baz$expected
moo
```

```
##        opinion
## color   dislike   like no opinion   so so
##   blue   35.638 17.098      25.75 24.514
##   green  81.656 39.176      59.00 56.168
##   red    55.706 26.726      40.25 38.318
```

These are the best estimates of the expected values *under the null hypothesis* (and we don't even know what that is yet).

For short, we call the observed data *observed*. The form of the Pearson chi-squared test statistic is

$$\sum_{\text{all cells}} \frac{(\text{observed} - \text{expected})^2}{\text{expected}} \tag{‡}$$

We didn't actually need to know this because R function `chisq.test` knows that.

```r
tstat <- sum((bar - moo)^2 / moo)
all.equal(tstat, baz$statistic, check.attributes = FALSE)
```

```
## [1] TRUE
```

The Pearson chi-squared test is a special case of the Rao (or score) test. There is also a likelihood ratio (or Wilks) test. It has a test statistic of this form.

$$\sum_{\text{all cells}} 2 \cdot \text{observed} \cdot \log \left( \frac{\text{observed}}{\text{expected}} \right) \tag{‡‡}$$

As always, these two tests are *asymptotically equivalent* (give nearly the same results for sufficiently large $n$, and here the sample size is large).

```r
sum(bar)
```

```
## [1] 500
```

So we use the same null distribution (distribution of the test statistic under the null hypothesis), which is the chi-squared distribution with

```r
baz$parameter
```

```
## df
##  6
```

degrees of freedom (not completely sure why it calls degrees of freedom "parameter", apparently so it can use the same print function for both tests about one parameter and tests of model comparison).

So the LRT (likelihood ratio test) $P$-value is

```r
tstat <- sum(2 * bar * log(bar / moo))
pchisq(tstat, df = baz$parameter, lower.tail = FALSE)
```

```
## [1] 0.0159272
```

So not much difference, $P = 0.018$ for the Pearson chi-squared test (special case of Rao test) and $P = 0.016$ for the likelihood ratio test.

**Warning:** the likelihood ratio test statistic calculation above blows up, giving `0 * -Inf = NaN` when any of the observed values are zero, but this is incorrect, just bad calculation. Since

$$x \log(x) \to 0, \qquad \text{as } x \to 0$$

the term in the sum for each cell having observed value zero should be zero, so to be careful we calculate

```r
tstat <- 2 * bar * log(bar / moo) # no sum yet
tstat[is.nan(tstat)] <- 0
tstat <- sum(tstat) # now OK to sum
```

## 6.4 R Function `glm` with `family = poisson`

So that was that. We are now done with how to do two tests (Pearson and LRT) for these data. But that is not how we are going to do higher dimensional tables. For that we are going to use R function `glm`. So we show how to do that too.

```
gout0 <- glm(y ~ color + opinion, family = poisson, data = foo)
gout1 <- glm(y ~ color * opinion, family = poisson, data = foo)
sout.lrt <- anova(gout0, gout1, test = "LRT")
sout.rao <- anova(gout0, gout1, test = "Rao")
```

Note that R function `glm` wants the data in data frame form.

```
sout.lrt
```

```
## Analysis of Deviance Table
##
## Model 1: y ~ color + opinion
## Model 2: y ~ color * opinion
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1         6     15.623
## 2         0      0.000  6   15.623  0.01593 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
sout.rao
```

```
## Analysis of Deviance Table
##
## Model 1: y ~ color + opinion
## Model 2: y ~ color * opinion
##   Resid. Df Resid. Dev Df Deviance    Rao Pr(>Chi)
## 1         6     15.623
## 2         0      0.000  6   15.623 15.371  0.01756 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
all.equal(sout.rao[2, "Pr(>Chi)"], baz$p.value)
```

```
## [1] "Mean relative difference: 7.453305e-06"
```

```
all.equal(sout.lrt[2, "Pr(>Chi)"],
    pchisq(tstat, df = baz$parameter, lower.tail = FALSE))
```

```
## [1] TRUE
```

```
all.equal(sout.rao[2, "Pr(>Chi)"], baz$p.value,
    tolerance = 1e-5)
```

```
## [1] TRUE
```

One or the other of these R functions is a bit sloppy about calculating the Rao test. No idea which. Some sloppiness is inherent in the fact that computer arithmetic is not exact (no more than 16 decimal place precision, but errors can add up in complicated calculations; also convergence can be sloppy in optimization algorithms  R function `chisq.test` does not do optimization but R function `glm` does).

## 6.5 Theory

### 6.5.1 Sampling Distributions for Hypothesis Tests

So now we get to theory. What are these hypothesis tests testing? And that depends on

- the family of distributions assumed for the data and
- the null hypothesis (which is in the assumed family).

And here is where it gets complicated *even in intro stats.* There are three different families of distributions that can be used

- Poisson,
- multinomial, and
- product multinomial.

But intro stats only distinguishes two

- tests of independence (of the categorical variables in the data frame, which are the row and column labels in the table) and
- tests of homogeneity of proportions (that the rows (or columns but not both) all have the same proportions in the same positions).

The correspondence between intro stats and sophisticated courses like this is

- tests of independence correspond to Poisson *and* multinomial and
- tests of homogeneity of proportions correspond to product multinomial.

So, first, a little bit about sampling models (families of distributions). More on that here and here.

- Poisson sampling is for when you go out and collect data for a fixed amount of time and the time at which you collect each data point and which cell of the table it goes in are unrelated to any other data point.
- multinomial sampling is like Poisson except the total sample size (sum of all counts) is fixed in advance and you continue sampling until you get that many and then stop.
- product multinomial sampling is like multinomial except the row sums or column sums but not both of the table are fixed in advance and you continue sampling until you get that many in each row (respectively, column) and then stop (in that row (respectively, column); you keep sampling the others until they get the required number).

So the difference is what is random and what is fixed.

- In Poisson, all counts are random.
- In multinomial, all counts are random subject to the sum being fixed.
- In product multinomial, all counts are random subject to the row (respectively column) sums being fixed.

And now comes the theoretical shocker. If we could do it (but we cannot) *exact* inference (based on *exact* sampling distributions of test statistics) would differ for the three sampling schemes. But the asymptotic approximations are *the same!* That is, whether we use Poisson, multinomial, or product multinomial for the sampling scheme (family of distributions) and whether we use the Wilks (also called likelihood ratio), Rao (also called score) or Wald test, we have (for large $n$)

- nearly the same test statistic,
- the same large sample approximate null distribution of the test statistic (chi-squared with some degrees of freedom, which is the same for all), and hence
- nearly the same (approximate, large $n$) $P$-value.

That is why all of our computing examples above *ignored the sampling scheme.* You get the same answer (using large sample approximation) for all three sampling schemes.

As with tests about binomial proportions, we did not illustrate Wald tests. This is because textbooks do not recommend them, even though they are asymptotically equivalent (work no better and *no worse* for large $n$) than the other two.

Also R function `anova` does not offer the Wald test as an option.

### 6.5.2 The Null Hypothesis.

#### 6.5.2.1 All Sampling Schemes

We are going to discuss the null hypothesis for *data in table form.* The notation would be different (discussed later in the course) for data frame form. But, of course, the data are the data, no matter how printed or stored in the computer. So data frame versus table make no difference to the statistical analysis.

Mathematically, the table is a matrix. The observed values are $y_{ij}$ where $i$ is the row index and $j$ is the column index, that is, $y_{ij}$ is the count in the cell in row $i$ and column $j$. The corresponding parameter is $\mu_{ij}$. These parameters are the mean values (or expected values) for the cells. The families of distributions are

- Poisson: the data in the cells of the table are independent Poisson random variables.
- Multinomial: the data in the cells of the table are components of a multinomial random vector.
- Product multinomial: the rows (or columns) of the table are independent multinomial random vectors.

More on what independent means in probability and statistics here. You should have some idea about that from previous statistics courses.

All three sampling distributions are determined by (the collection of) means $\mu_{ij}$. Thus those are parameters for all the models.

The null hypothesis uses the multiplicative assumption

$$\mu_{ij} = \alpha_i \beta_j, \qquad \text{for all } i \text{ and } j,$$

where the $\alpha_i$ and $\beta_j$ are the parameters of the family of distributions that is the null hypothesis.

The reason the tests discussed above are called tests of independence is whenever means are products, that comes from independence.

There is something a bit funny about this parameterization in that it is not *identifiable.* (You may have seen this topic discussed in other statistics courses using the term *collinearity* or multicollinearity.) If we divide all of the $\alpha_i$ by a constant $c$ and multiply all the $\beta_j$ by the same constant we get the same means

$$\mu_{ij} = \alpha_i \beta_j = \frac{\alpha_i}{c} \cdot c\beta_j \tag{§}$$

So the $\alpha_i$ and $\beta_j$ uniquely determine the $\mu_{ij}$ but *not vice versa.* Thus we cannot estimate all of the alphas and all of the betas: they are not *identifiable.* But we can fix any one of the alphas or betas at an arbitrary positive value, and then estimate the rest (this just amounts to choosing a particular constant $c$).

#### 6.5.2.2 Multinomial Sampling

When we are assuming multinomial sampling, we often use a different parameterization

$$\mu_{ij} = n\pi_{ij}$$

so

$$\pi_{ij} = \frac{\mu_{ij}}{n}$$

where (as usual) $n$ is the sample size, the total number of counts in the whole table, the total number of individuals surveyed.

Recall that in multinomial sampling $n$ is not a parameter. It is a number fixed in advance of any data being collected. It is a known constant, unlike the parameters which are *unknown* constants.

Also probabilities sum to one, so we must have

$$\sum_{ij} \pi_{ij} = 1 \tag{1}$$

$$\sum_{ij} \mu_{ij} = n \tag{2}$$

Of course since the $\pi_{ij}$ determine the $\mu_{ij}$ and vice versa, they are equivalent; they parameterize the same model. And if the $\mu_{ij}$ have the multiplicative form, then so do the $\pi_{ij}$ and vice versa. But for these we sometimes redefine the alphas and betas, writing

$$\pi_{ij} = \alpha_i \beta_j$$

so

$$\mu_{ij} = n \alpha_i \beta_j$$

Now we have the restriction that probabilities must sum to one, so

$$\sum_{ij} \pi_{ij} = 1 = \left(\sum_i \alpha_i\right) \left(\sum_j \beta_j\right)$$

and we see that one way to obtain this is to have the $\alpha_i$ to sum to one and the $\beta_j$ to sum to one. We don't have to do this. We only need the $\pi_{ij}$ to sum to one. But this is convenient and makes $\alpha$ and $\beta$ probability vectors (have components that are nonnegative and sum to one).

Introduce some convenient notation

$$y_{i+} = \sum_j y_{ij}$$

$$y_{+j} = \sum_i y_{ij}$$

$$y_{++} = \sum_{ij} y_{ij}$$

then the best estimates of the parameters of the null hypothesis are

$$\hat{\alpha}_i = \frac{y_{i+}}{y_{++}} \tag{3}$$

$$\hat{\beta}_j = \frac{y_{+j}}{y_{++}} \tag{4}$$

$$\hat{\pi}_{ij} = \hat{\alpha}_i \hat{\beta}_j \tag{5}$$

$$\hat{\mu}_{ij} = n \hat{\alpha}_i \hat{\beta}_j \tag{6}$$

We have hats on everything to remind us that these are just parameter estimates (not the true unknown parameter values).

#### 6.5.2.3 Poisson Sampling

The same numbers are the parameter estimates for Poisson sampling. If we want to not mention the $\pi_{ij}$ because they only make sense for multinomial sampling, then we can plug earlier equations into the last equation to get

$$\hat{\mu}_{ij} = \frac{y_{i+} y_{+j}}{y_{++}} \tag{§§}$$

In either case our formulas agree with the output of R functions `chisq.test` and `glm`.

```
mu.hat <- outer(rowSums(bar), colSums(bar)) / sum(bar)
all.equal(mu.hat, baz$expected, check.attributes = FALSE)
```

```
## [1] TRUE
```

```
qux <- transform(foo, mu.hat = gout0$fitted.values)
moo <- xtabs(mu.hat ~ color + opinion, data = qux)
moo <- unclass(moo)
all.equal(mu.hat, moo, check.attributes = FALSE)
```

```
## [1] TRUE
```

Sometimes it can be a real pain to get R to check that two different calculations give the same results (except for inaccuracy of computer arithmetic). Here the means produced by R function `glm` are a vector. So we had to use R function `xtabs` to put the means in the right cells of the contingency table. And we had to use R function `unclass` to get rid of the class attribute for R object `moo` because otherwise R function `all.equal` says one is a `"matrix"` and the other `"xtabs"` and won't do the comparison.

#### 6.5.2.4 Product Multinomial Sampling

Now we still have multiplicative form of the null hypothesis and (§§) still gives the best estimates of the parameters under the null hypothesis. But now we write the null hypothesis

$$\mu_{ij} = n_i \beta_j \tag{§§§}$$

if we are assuming that row sums of the data $(n_i)$ were fixed in advance and

$$\mu_{ij} = n_j \alpha_i$$

if we are assuming that column sums of the data $(n_j)$ were fixed in advance. Now our only parameters are $\beta_j$ in the first and $\alpha_i$ in the second.

And we say the null hypothesis specifies *homogeneity of proportions*, that is, the same vector $\beta$ with components $\beta_j$ applies to each row of the table in the first case, and, the same vector $\alpha$ with components $\alpha_i$ applies to each column of the table in the first case,

All of this seems complicated, but for all the sampling schemes the null hypothesis has the form

$$\text{expected}_{ij} = (\text{something})_i (\text{something else})_j$$

regardless of what notation we use, and regardless of what parts are unknown parameters and what parts are known sample sizes.

Also for all three sampling schemes

- the likelihood ratio test statistic is the same,
- the (approximate, large $n$) null distribution of the test statistic is the same, and hence
- the (approximate, large $n$) $P$-value is the same

Of course this is only for the different sampling schemes. For the different test statistics (Wilks, Rao, Wald), repeating what was said above, for large $n$,

- the test statistics are nearly the same,
- the approximate null distribution of the test statistics are exactly the same, and hence
- the approximate $P$-values are nearly the same.

### 6.5.3 The Alternative Hypothesis

For all of these procedures, the alternative hypothesis is "anything goes", that is the parameters can be anything other than in the null hypothesis.

### 6.5.4 Degrees of Freedom

#### 6.5.4.1 General Rule

The general rule for degrees of freedom for Wilks, Rao, and Wald tests is: degrees of freedom is the number of identifiable parameters in the alternative hypothesis minus the number of identifiable parameters in null hypothesis.

Here *identifiable* means different identifiable parameter vectors correspond to different distributions. There are no constraints on their values. In multinomial and product multinomial we do have some constraints. And those constraints change the numbers of identifiable parameters.

#### 6.5.4.2 Poisson

Suppose our table has $r$ rows and $c$ columns. Then the $\mu_{ij}$ can be any nonnegative numbers. So

- the alternative hypothesis has $rc$ identifiable parameters, and
- the null hypothesis has $r + c - 1$ identifiable parameters.

The latter is because the adjustable constant $c$ in (§) can be chosen arbitrarily, and this makes one of the alphas (or one of the betas but not both) a known constant (rather than an unknown parameter to be estimated). For example, if we choose $c = \alpha_1$, then the new $\alpha$ vector (the old $\alpha$ vector divided by $c$) has first component $\alpha_1 = 1$.

Hence the degrees of freedom is

$$(rc) - (r + c - 1) = (r - 1)(c - 1)$$

#### 6.5.4.3 Multinomial

Now

- the alternative hypothesis has $rc - 1$ identifiable parameters, and
- the null hypothesis has $r + c - 2$ identifiable parameters.

The reason for the first is the constraint that the $\pi_{ij}$ must sum to one, so any single $\pi_{ij}$ is determined by the rest.

The reason for the second is the constraints that the $\alpha_i$ must sum to one, and so must the $\beta_j$. Hence there are $r - 1$ identifiable alphas and $c - 1$ identifiable betas.

Hence the degrees of freedom is

$$(rc - 1) - (r + c - 2) = (r - 1)(c - 1)$$

#### 6.5.4.4 Product Multinomial

We do the case where row sums are fixed in advance so the null hypothesis is (§§§). The other case is similar. Now

- the alternative hypothesis has $r(c - 1)$ identifiable parameters, and
- the null hypothesis has $c - 1$ identifiable parameters.

The latter is because the $\beta_j$ must sum to one, so any single $\beta_j$ is determined by the rest.

The alternative hypothesis has different $\beta$ vectors for different rows, and each has $c-1$ identifiable parameters. So $r(c - 1)$.

Hence the degrees of freedom is

$$r(c - 1) - (c - 1) = (r - 1)(c - 1)$$

We waded through all of this algebra to give a concrete example. But we needn't have. It is a theorem that Wilks, Rao, and Wald tests comparing the same hypotheses always have the same asymptotic distribution, which is chi-squared with the same degrees of freedom. And it is another theorem that the degrees of freedom do not depend on the sampling scheme; it is the same for Poisson, multinomial, or product multinomial sampling.

So once we have done one (correct) degrees of freedom calculation, we know all the others have to agree *by theory*.

Also, R functions can usually figure out the degrees of freedom for us. We only need to figure it out by theory when there is no available R function to do the analysis.

# 7 Interpretation

A standard thing to ask students on homeworks or tests is to "interpret" your results, that is, not just give some numbers or pictures but rather to explain what they mean.

## 7.1 Bayesian

Bayesian inference almost needs no interpretation. A posterior distribution or a prior distribution is a probability distribution, and if someone doesn't know what that is, then they don't have enough education to understand Bayesian inference (here are our notes on that).

One point is that you must make clear what parameterization is being used. We have already seen that there can be more than one parameter used to describe a statistical model. So you have to be clear about which one.

## 7.2 Frequentist Confidence Intervals

This is even more essential for confidence intervals. A confidence interval is an interval estimate of a *parameter*. If you don't say which parameter, then your interval is useless. In our example of fixing up a Wald interval we had an interval for $\pi$

```
(pout$fit + c(-1, 1) * crit * pout$se.fit) |> invlogit()
```

```
## [1] 0.4592938 0.9495868
```

and this corresponds to an interval for $\theta$

```
(pout$fit + c(-1, 1) * crit * pout$se.fit)
```

```
## [1] -0.163186  2.935775
```

Clearly, if you don't say which interval is for which parameter, what you are saying is meaningless.

Other points to make about interpreting confidence intervals are just facts about frequentist inference

- the confidence interval may not cover the true unknown parameter value (it's called a 95% confidence interval because it misses 5% of the time),
- unlike Bayesians, we are not treating the parameter as random, what is random is the endpoints of the confidence interval, and
- you should not refer to the parameter as some Greek letter but rather something your clients or readers can understand: say "the proportion of Minnesotans whose favorite ice cream flavor is mint chocolate chip" (or whatever the parameter is).

## 7.3 Frequentist Hypothesis Tests

A $P$-value interprets itself. That is why the concept was invented. A $P$-value is not made more understandable by declaring that it means “statistically significant” results or attaching other adjectives, like “borderline statistically significant” or some such.

People *want* clear cut answers. Your theory is either proved or not. But statistics *does not give clear cut answers.* Trying to make it seem like it does is always wrong.

$P$-values are numbers between zero and one. Low $P$-values (near zero) are evidence against the null hypothesis, the lower the $P$-value, the stronger the inference. High $P$-values (near one) are still evidence against the null hypothesis, but the higher the $P$-value, the weaker the inference. For $P$-values above 0.1 the evidence against the null hypothesis is so weak as to be unimpressive. Even more so for higher $P$-values.

Some intro stats books outlaw the phrase “accept the null hypothesis” (appropriate when the $P$-value is large) and insist on the phrase “fail to reject the null hypothesis”.

The point is, that the $P$-value is only relevant to the data used in the analysis. A large $P$-value only says that *these data* do not show any “statistically significant” lack of fit of the null hypothesis. More data (if you were to get some) might show lack of fit (or might not).

The traditional dividing line between $P$-values being “low” or “high” is 0.05. But this tradition has nothing to recommend it. It is just laziness on the part of those using it. True, 0.05 is a nice “round” number, but it is only a round number if we use the decimal system, which we consider natural because people have ten fingers. So using 0.05 as a cutoff for $P$-values is trying to decide important scientific (business, sports, whatever) questions by counting on your fingers.

I once wrote

> Anyone who thinks there is an important difference between $P = 0.049$ and $P = 0.051$ understands neither science nor statistics.

But my coauthor cut it — not because it was wrong but because it might offend.

Don’t you ever seem to run afoul of this dictum. Never say anything that can be interpreted as saying there is an important difference between $P = 0.049$ and $P = 0.051$.

$P$-values fall on a continuum, and no point on that continuum is special.

As to what $P$-values are, the answer is tricky.

When we have a point null hypothesis (that completely specifies a probability model), then the $P$-value is a probability. It is the probability *assuming the null hypothesis is true* of seeing a result that gives at least as much *apparent* evidence *against* the null hypothesis. In short, it is the probability of a false positive, of *falsely* claiming to have statistically significant results.

When we are doing approximate rather than exact inference  using large sample approximation, as we are always doing in this course (except for fuzzy)  then a $P$-value is the same except for approximately. It is *approximately* the probability of a false positive.

So the lower the $P$-value, the lower the probability of *falsely* claiming to have discovered something.

But that probability of getting the wrong answer is never zero.

When we have a composite null hypothesis, for example,

$$\begin{aligned} H_0 &: \text{true unknown parameter value} \geq \theta \\ H_1 &: \text{true unknown parameter value} < \theta \end{aligned}$$

for a one-tailed test, a $P$-value is not a probability but rather an upper bound on probabilities, but we won’t worry about that. This does not apply to approximate (large sample) tests. For those a $P$-value is still an approximate probability.

One last thing about hypothesis tests. If you don't mention what the null hypothesis is mathematically and what (if anything) this means in scientific (business, sports, whatever) terms, then you haven't explained anything about the hypothesis test. It is all about the null hypothesis. So if you don't understand the null hypothesis, then you don't understand anything about the hypothesis test.

For example, for the null hypothesis of independence discussed above, one should say which random variables, *color* and *opinion*, have been found to be dependent ($P = 0.0159$).

One should also be very careful to avoid overclaiming what the test says.

- We all know *correlation is not causation* and *stochastic dependence isn't either*. So don't say that!
- Some dependence does not mean that any particular theory (causal or not) about the particular form of the dependence is true. Rejecting the null hypothesis means there is *some* dependence of *some sort or other*, not necessarily any particular form.