Chi-squared (χ^2) (1.10.5) and F-tests (9.5.2) for the variance of a normal distribution

χ^2 tests for goodness of fit and indepdendence (3.5.4–3.5.5)

Prof. Tesler

Math 283
Fall 2016

Tests of means vs. tests of variances

Data $x_1, \ldots, x_n$, sample mean $\bar{x}$, sample var. ${s_X}^2$
Data $y_1, \ldots, y_m$, sample mean $\bar{y}$, sample var. ${s_Y}^2$

Tests for mean

One-sample tests:

$H_0 : \mu = \mu_0$ vs. $H_1 : \mu \neq \mu_0$

test statistic:

$z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}$ or $t = \frac{\bar{x} - \mu_0}{s / \sqrt{n}}$ $(df = n-1)$

Two-sample tests:

$H_0 : \mu_X = \mu_Y$ vs. $H_1 : \mu_X \neq \mu_Y$

test statistic:

$$z = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{{\sigma_X}^2}{n} + \frac{{\sigma_Y}^2}{m}}}$$

or $t = \frac{\bar{x} - \bar{y}}{s_p \sqrt{\frac{1}{n} + \frac{1}{m}}}$ $(df = n + m - 2)$

Tests for variance

One-sample test:

$H_0 : \sigma^2 = {\sigma_0}^2$ vs. $H_1 : \sigma^2 \neq {\sigma_0}^2$

test statistic: "chi-squared"

$\chi^2 = (n-1)s^2 / {\sigma_0}^2$ $(df = n-1)$

Two-sample test:

$H_0 : {\sigma_X}^2 = {\sigma_Y}^2$ vs. $H_1 : {\sigma_X}^2 \neq {\sigma_Y}^2$

test statistic:

$F = {s_Y}^2 / {s_X}^2$
(with $m-1$ and $n-1$ d.f.)

Application: The fine print in the Z and t-tests

- **One-sample z-test,** H_0: $\mu = \mu_0$ vs. H_1: $\mu \neq \mu_0$
  - This assumes that you know the value of σ^2, say $\sigma^2 = {\sigma_0}^2$.
  - A χ^2 test could be used to verify that the data is consistent with $H_0 : \sigma^2 = {\sigma_0}^2$ instead of $H_1 : \sigma^2 \neq {\sigma_0}^2$.
- **Two-sample z-test,** H_0: $\mu_X = \mu_Y$ vs. H_1: $\mu_X \neq \mu_Y$
  - This assumes that you know the values of ${\sigma_X}^2$ and ${\sigma_Y}^2$.
  - Separate χ^2 tests for ${\sigma_X}^2$ and ${\sigma_Y}^2$ could be performed to verify consistency with the assumed values.
- **Two-sample t-test,** H_0: $\mu_X = \mu_Y$ vs. H_1: $\mu_X \neq \mu_Y$
  - This assumes ${\sigma_X}^2 = {\sigma_Y}^2$ (but doesn't assume that this common value is known to you).
  - An F-test could be used to verify that the data is consistent with $H_0 : {\sigma_X}^2 = {\sigma_Y}^2$ instead of $H_1 : {\sigma_X}^2 \neq {\sigma_Y}^2$.
  - If the variances are unequal, Welch's t-test can be used instead of the regular two-sample t-test (Ewens & Grant pp. 127–128).

The χ^2 ("Chi-squared") distribution

- Used for confidence intervals and hypothesis tests on the unknown parameter σ^2 of the normal distribution, based on the test statistic s^2 (sample variance).
- It has the same "degrees of freedom" as for the t distribution.

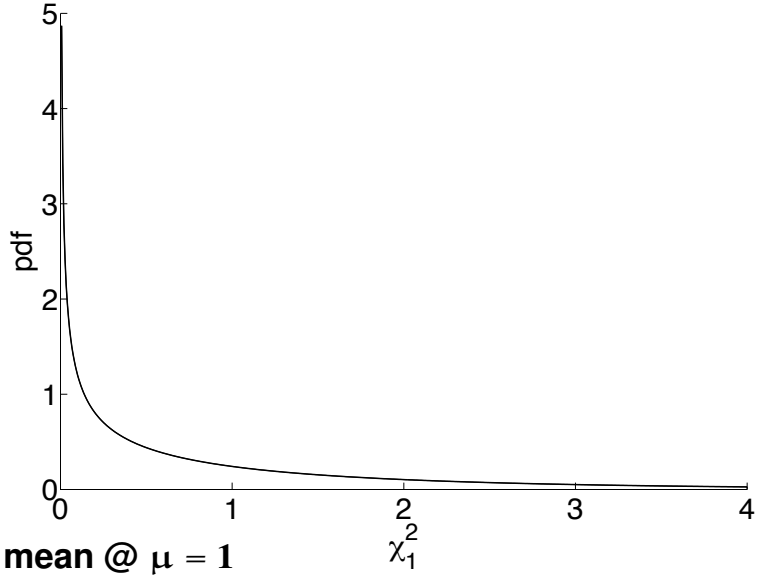
Point these out on the graphs:

- The chi-squared distribution with k degrees of freedom has

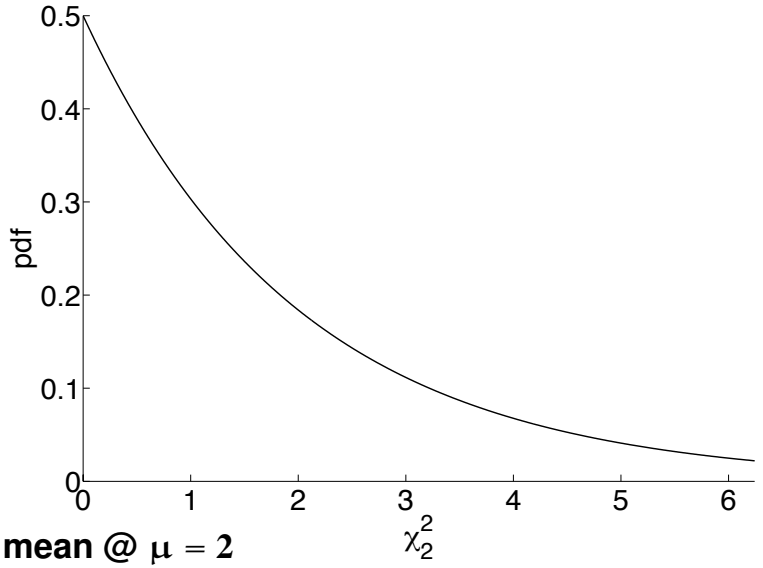
Range	$[0, \infty)$
Mean	$\mu = k$
Mode	$\chi^2 = k-2$ (for $k \geqslant 2$, the pdf is maximum for $\chi^2 = k-2$) $\chi^2 = 1$ (for $k = 1$)
Median	$\approx k(1 - \frac{2}{9k})^3$ Between k and $k - \frac{2}{3}$. Asymptotically decreases $\to k - \frac{2}{3}$ as $k \to \infty$.
Variance	$\sigma^2 = 2k$
PDF	$\frac{x^{(k/2)-1}e^{-x/2}}{2^{k/2}\Gamma(k/2)}$: Γ distrib. with shape $r = \frac{k}{2}$, rate $\lambda = \frac{1}{2}$

- Unlike z and t, the pdf for χ^2 is NOT symmetric.

The graphs for 1 and 2 degrees of freedom are decreasing:

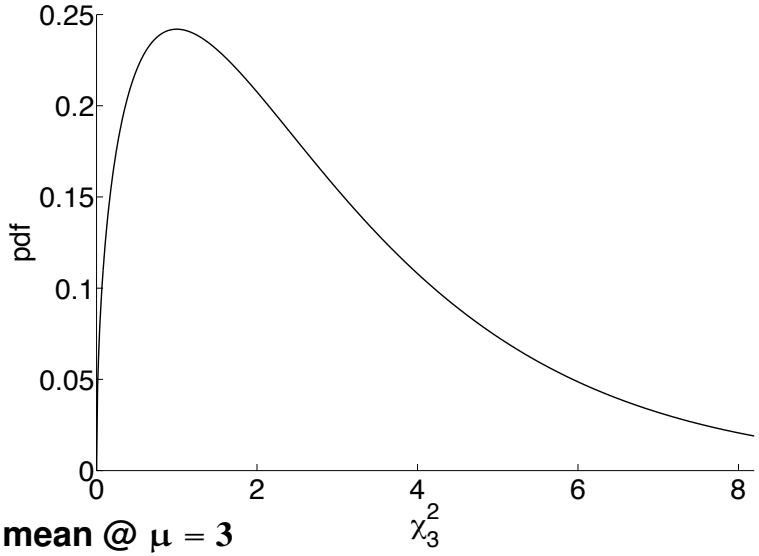


mean @ $\mu = 1$
mode @ $\chi^2 = 0$
median @ χ^2 = `chi2inv(.5,1)` = `qchisq(.5,1)` **= 0.4549**

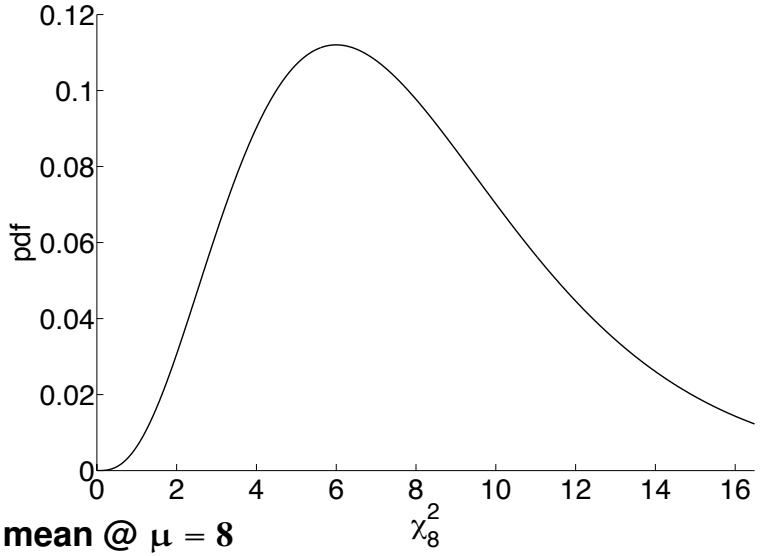


mean @ $\mu = 2$
mode @ $\chi^2 = 0$
median @ χ^2 = `chi2inv(.5,2)` = `qchisq(.5,2)` **= 1.3863**

The rest are "hump" shaped and skewed to the right:

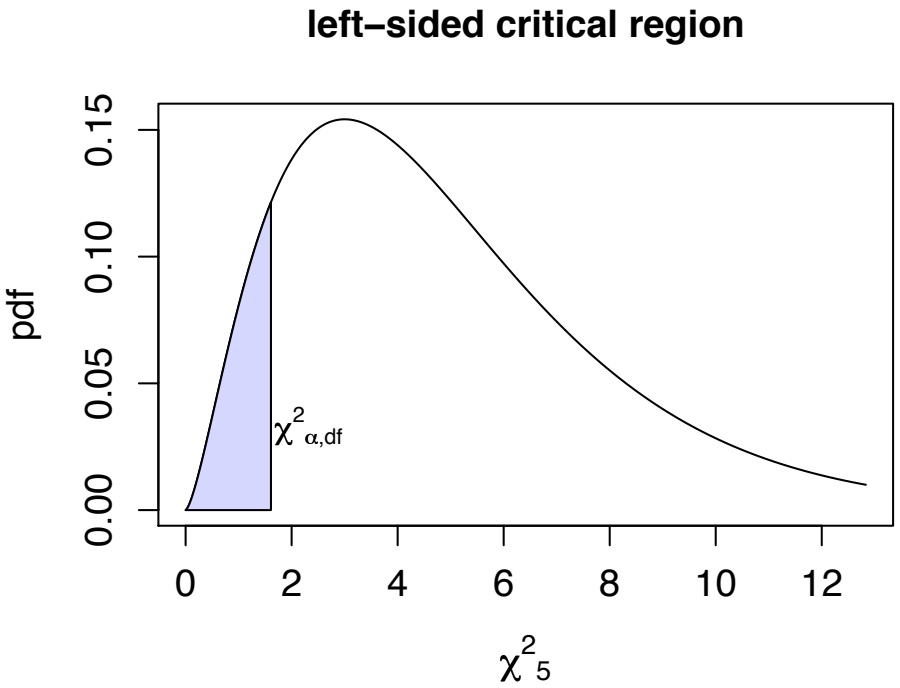


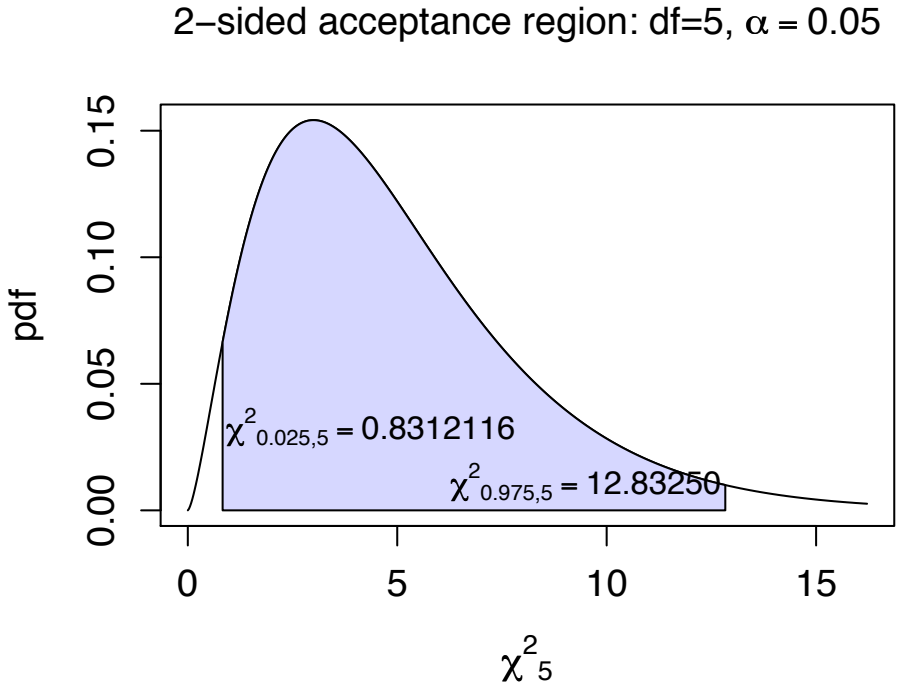
mean @ $\mu = 3$
mode @ $\chi^2 = 1$
median @ χ^2 = `chi2inv(.5,3)` = `qchisq(.5,3)` **= 2.3660**



mean @ $\mu = 8$
mode @ $\chi^2 = 6$
median @ χ^2 = `chi2inv(.5,8)` = `qchisq(.5,8)` **= 7.3441**

χ^2 ("Chi-squared") distribution — Cutoffs

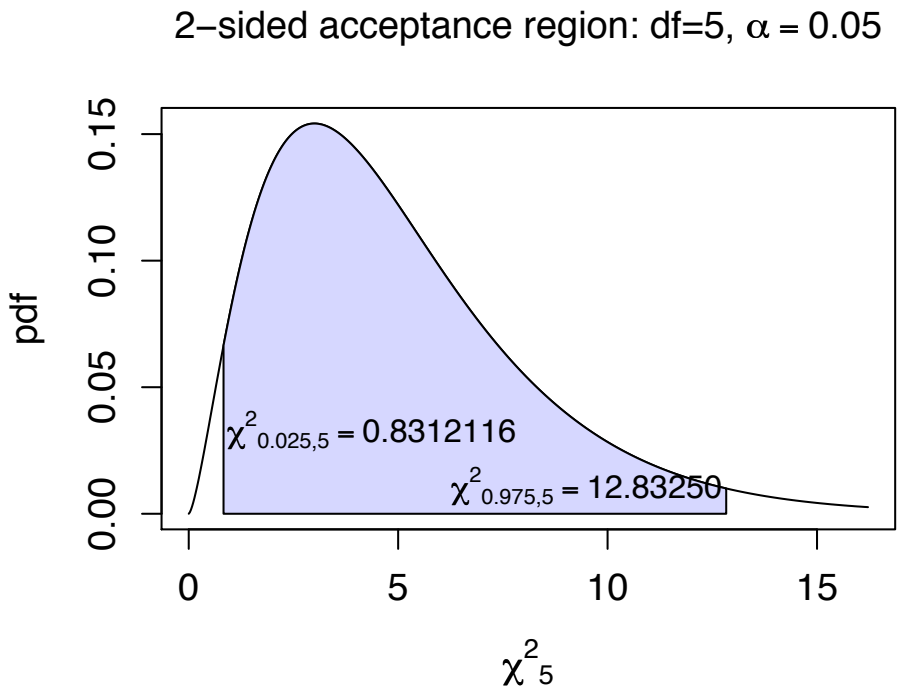




- Define $\chi^2_{\alpha,df}$ as the number where the cdf (area *left* of it) is α: $P(\chi^2_{df} \leqslant \chi^2_{\alpha,df}) = \alpha$
- Different notation than z_α and $t_{\alpha,df}$ (area α on *right*) since pdf isn't symmetric.

	Matlab		**R**	
$\chi^2_{0.025,5}$	`chi2inv(.025,5)`	=	`qchisq(.025,5)`	$= 0.8312$
$\chi^2_{0.975,5}$	`chi2inv(.975,5)`	=	`qchisq(.975,5)`	$= 12.8325$
	`chi2cdf(0.8312,5)`	=	`pchisq(0.8312,5)`	$= 0.025$
	`chi2cdf(12.8325,5)`	=	`pchisq(12.8325,5)`	$= 0.975$
	`chi2pdf(0.8312,5)`	=	`dchisq(0.8312,5)`	$= 0.0665$
	`chi2pdf(12.8325,5)`	=	`dchisq(12.8325,5)`	$= 0.0100$

Two-sided cutoff



- The mean, median, and mode are different, so it may not be obvious what values of χ^2 are "more consistent" with the null H_0: $\sigma^2 = 10000$ vs. the alternative $\sigma^2 \neq 10000$.
- **Closer to the median of χ^2 is "more consistent" with H_0.**
- For 2-sided hypothesis tests or confidence intervals with $\alpha = 5\%$, we still put 95% of the area in the middle and 2.5% at each end, but the pdf is not symmetric, so the lower and upper cutoffs are determined separately instead of $\pm$ each other.

Two-sided hypothesis test for variance

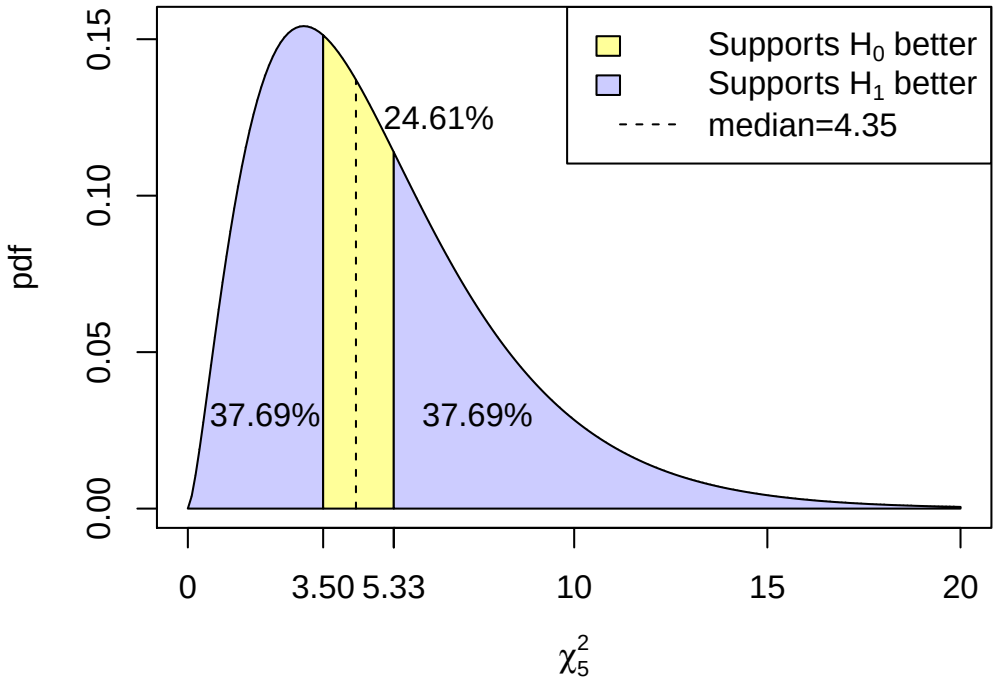
Test $H_0 : \sigma^2 = 10000$ vs. $H_1 : \sigma^2 \neq 10000$ at sig. level $\alpha = .05$
(In general, replace 10000 by ${\sigma_0}^2$; here, $\sigma_0 = 100$)

Decision procedure

1. Get a sample $x_1, \ldots, x_n$.
   650, 510, 470, 570, 410, 370 with $n = 6$
2. Calculate $m = \frac{x_1 + \cdots + x_n}{n}$ and $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - m)^2$.
   $m = 496.67, \quad s^2 = 10666.67, \quad s = 103.28$
3. Calculate the test-statistic $\chi^2 = \frac{(n-1)s^2}{{\sigma_0}^2} = \sum_{i=1}^{n}\frac{(x_i - m)^2}{{\sigma_0}^2}$
   $\chi^2 = \frac{(n-1)s^2}{{\sigma_0}^2} = \frac{(6-1)(10666.67)}{10000} = 5.33$
4. Accept H_0 if χ^2 is between $\chi^2_{\alpha/2,n-1}$ and $\chi^2_{1-\alpha/2,n-1}$.
   Reject H_0 otherwise.
   $\chi^2_{.025,5} = .8312, \quad \chi^2_{.975,5} = 12.8325$

   Since $\chi^2 = 5.33$ is between these, we accept H_0.
   (Or, there is insufficient evidence to reject $\sigma^2 = 10000$.)

Doing the same test with a P-value



- $P(\chi^2_5 \leqslant 5.33) = 0.6231$ is the area left of 5.33 for χ^2 with 5 d.f.:
  Matlab: `chi2cdf(5.33,5)` **R:** `pchisq(5.33,5)`
- Values at least as extreme as this are those at the 62.31th percentile or higher, OR at the 37.69th percentile or lower, so
$$P = (1 - .6231) + .3769 = 2(.3769) = 0.7539$$
- $P > \alpha$ $(0.75 > 0.05)$ so accept H_0.
- To turn a one-sided P-value p_1 into a two-sided P-value, use
$$P = 2\min(p_1, 1 - p_1).$$

Two-sided 95% confidence interval for the variance

Continue with data 650, 510, 470, 570, 410, 370
which has $n = 6, \quad m = 496.67, \quad s^2 = 10666.67, \quad s = 103.28.$

- Get bounds on σ^2 in terms of s^2 for the two-sided test:

$$\begin{aligned} 0.95 &= P(\chi^2_{.025,5} < \chi^2 < \chi^2_{.975,5}) \\ &= P(0.8312 < \chi^2 < 12.8325) \\ &= P\left(0.8312 < \frac{(6-1)S^2}{\sigma^2} < 12.8325\right) \\ &= P\left(\frac{(6-1)S^2}{0.8312} > \sigma^2 > \frac{(6-1)S^2}{12.8325}\right) \end{aligned}$$

- A two-sided 95% confidence interval for the variance σ^2 is

$$\left(\frac{(6-1)S^2}{12.8325}, \frac{(6-1)S^2}{0.8312}\right) = (4156.11, 64164.26)$$

- A two-sided 95% confidence interval for σ is

$$\left(\sqrt{\frac{(6-1)S^2}{12.8325}}, \sqrt{\frac{(6-1)S^2}{0.8312}}\right) = (64.47, 253.31)$$

Properties of Chi-squared distribution

1. **Definition of Chi-squared distribution:**

   Let $Z_1, \ldots, Z_k$ be independent standard normal variables.

   Let $\chi^2_k = {Z_1}^2 + \cdots + {Z_k}^2$.

   The pdf of the random variable χ^2_k is the "chi-squared distribution with k degrees of freedom."

2. **Pooling property:** If U and V are independent χ^2 random variables with q and r degrees of freedom respectively, then $U + V$ is a χ^2 random variable with $q + r$ degrees of freedom.

3. **Sample variance:** Pick $X_1, \ldots, X_n$ from a normal distribution $N(\mu, \sigma^2)$. It turns out that

$$\sum_{i=1}^{n} \frac{(X_i - \overline{X})^2}{\sigma^2} = \frac{SS}{\sigma^2} = \frac{(n-1)S^2}{\sigma^2}$$

   has a χ^2 distribution with $df = n - 1$, so we test on $\chi^2 = \frac{(n-1)s^2}{{\sigma_0}^2}$.

Distributions with an additive/pooling property

For certain families of distributions, if U, V are independent random variables of that type, then $U + V$ is too, with certain parameters combining additively.

	Parameters of		
Distribution	U	V	$U + V$
Binomial	(n, p)	(m, p)	$(n + m, p)$
Negative binomial	(r, p)	(s, p)	$(r + s, p)$
Gamma	(r, λ)	(s, λ)	$(r + s, \lambda)$
Poisson	μ	ν	$\mu + \nu$
χ^2	q d.f.	r d.f.	$q + r$ d.f.

F distribution

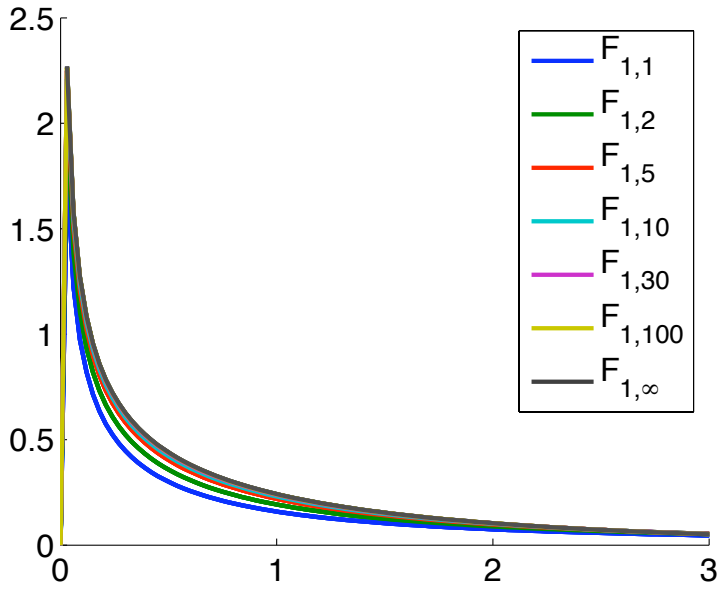
Let U and V be independent χ^2 random variables with q and r degrees of freedom, respectively. Then the random variable

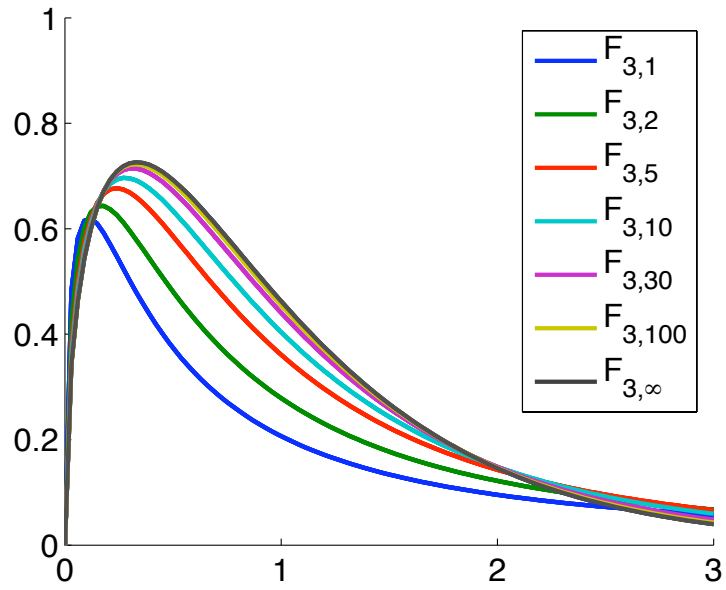
$$F = F_{q,r} = \frac{U/q}{V/r}$$

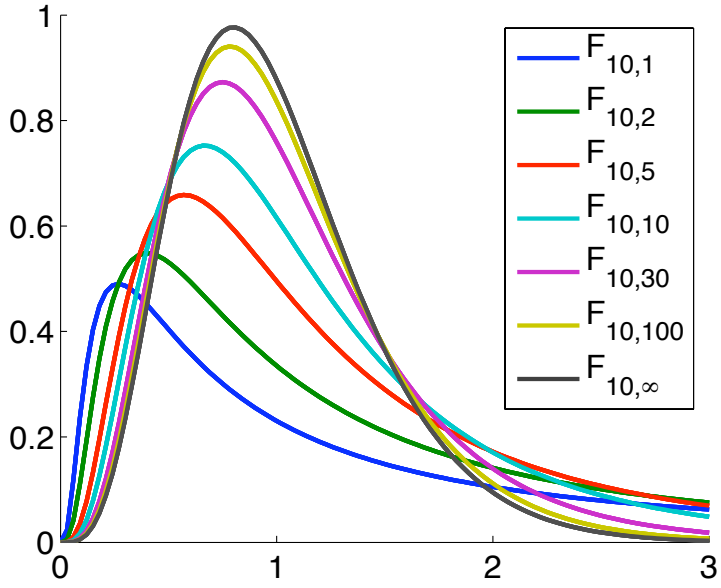
is called the "F distribution with q and r degrees of freedom."

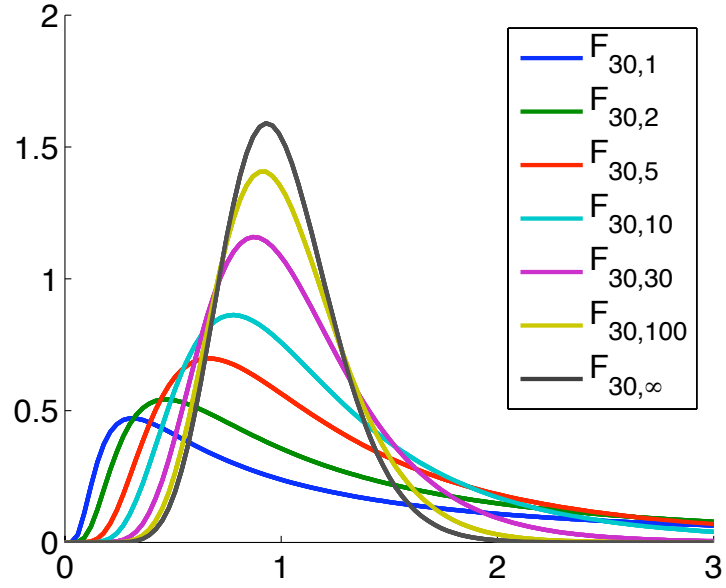
Range	$[0, \infty)$	**Variance**	$\frac{2r^2(q+r-2)}{q(r-2)^2(r-4)}$ if $r > 4$
Mean	$\frac{r}{r-2}$ if $r > 2$	**PDF**	messy
Mode	$\frac{r(q-2)}{q(r+2)}$ if $q > 2$		
Median	1 if $q = r$ > 1 if $q > r$ < 1 if $q < r$		

F distribution



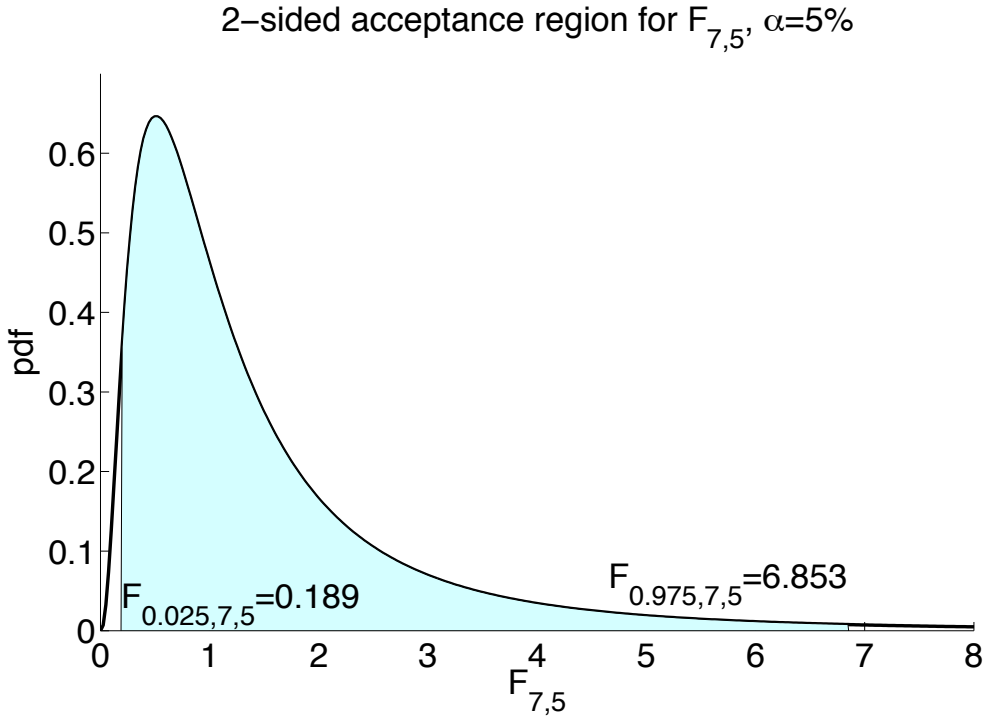






F distribution

Define $F_{\alpha,q,r}$ as the number where $P(F \leqslant F_{\alpha,q,r}) = \alpha$ (*left*-hand area α)



	Matlab		R	
$F_{0.025,7,5} =$	`finv(.025,7,5)`	$=$	`qf(.025,7,5)`	$= 0.1892$
$F_{0.975,7,5} =$	`finv(.975,7,5)`	$=$	`qf(.975,7,5)`	$= 6.853$
	`fcdf(0.1892,7,5)`	$=$	`pf(0.1892,7,5)`	$= 0.025$
	`fcdf(6.853,7,5)`	$=$	`pf(6.853,7,5)`	$= 0.975$
	`fpdf(0.1892,7,5)`	$=$	`df(0.1892,7,5)`	$= 0.3353$
	`fpdf(6.853,7,5)`	$=$	`df(6.853,7,5)`	$= 0.0077$

R's command `df` *is probability density of the* F *distribution,* **not** *degrees of freedom.*

Two-sided hypothesis test for F

- **Test at significance level $\alpha = 5\%$:**
  $H_0 : {\sigma_X}^2 = {\sigma_Y}^2$ vs. $H_1 : {\sigma_X}^2 \neq {\sigma_Y}^2$

- **Data for X:** $650, 510, 470, 570, 410, 370$
  $\bar{x} = 496.67, \quad {s_X}^2 = 10666.67, \quad s_X = 103.28, \quad df = 6 - 1 = 5$

- **Data for Y:** 510, 420, 520, 360, 470, 530, 550, 490
  $\bar{y} = 481.25, \quad {s_Y}^2 = 4012.5, \quad s_Y = 63.3443, \quad df = 8 - 1 = 7$

- **Test statistic:** $F = F_{7,5} = {s_Y}^2/{s_X}^2 = \frac{4012.5}{10666.67} = 0.3762.$

- Since our test statistic 0.3762 lies between the cutoffs $F_{.025,7,5} = 0.1892$ and $F_{.975,7,5} = 6.8531$, we accept H_0 / reject H_1.

P-values

- The CDF is $P(F_{7,5} \leqslant 0.3762) = 0.1174$

  Matlab: `fcdf(.3762,7,5)`

  R: `pf(.3762,7,5)`

- To make it two sided,

$$P = 2\min(0.1174, 1 - 0.1174) = 2(0.1174) = 0.2348$$

- Since $P > \alpha$ $(0.2348 > 0.05)$, we accept the null hypothesis.

F statistic to compare variances (two-sample data)

Theoretical setup for $H_0\colon \sigma_X^2 = \sigma_Y^2$ vs. $H_1\colon \sigma_X^2 \neq \sigma_Y^2$

- **First sample**: $x_1, \ldots, x_n$

$$V = \frac{(n-1)s_X^2}{\sigma_X^2} = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{\sigma_X^2} \quad \text{with } n-1 \text{ d.f.}$$

- **Second sample:** $y_1, \ldots, y_m$

$$U = \frac{(m-1)s_Y^2}{\sigma_Y^2} = \sum_{j=1}^{m} \frac{(y_i - \bar{y})^2}{\sigma_Y^2} \quad \text{with } m-1 \text{ d.f.}$$

- Assuming the null hypothesis $\sigma_X^2 = \sigma_Y^2$, the variances cancel:

$$F = \frac{U/(m-1)}{V/(n-1)} = \frac{s_Y^2/\sigma_Y^2}{s_X^2/\sigma_X^2} = \frac{s_Y^2}{s_X^2} \quad \text{with } m-1 \text{ and } n-1 \text{ d.f.}$$

- For

$$H_0\colon \sigma_X^2 = C\sigma_Y^2 \quad \text{vs.} \quad H_1\colon \sigma_X^2 \neq C\sigma_Y^2$$

where $C > 0$ is constant, use $F = Cs_Y^2/s_X^2$ instead.

k-sample experiments

- ANOVA (Analysis of Variance) is a procedure to compare the *means* of k-sample data (analagous to two-sample data, but for k independent sets of data).

- It involves the F distribution with a formula for F that takes into account all k samples instead of just two samples.

χ^2 tests for goodness of fit and independence (3.5.4–3.5.5)

Multinomial test

- Consider a k-sided die with faces $1, 2, \ldots, k$.
- We want to simultaneously test that the probabilities $p_1, p_2, \ldots, p_k$ of rolling $1, 2, \ldots, k$ are specified values.
- To test if a 6-sided die is fair,
  H_0: $(p_1, \ldots, p_6) = (1/6, \ldots, 1/6)$
  H_1: At least one $p_i \neq 1/6$
- Decision rule is based counting # 1's, 2's, etc. on n independent rolls of the die.
- For the fair coin problem, the exact distribution was binomial, and we approximated it with a normal distribution.
- For this problem, the exact distribution is multinomial.
  We will combine the separate counts of $1, 2, \ldots$ into a single test statistic whose distribution is approximately a χ^2 distribution.

Goodness of fit tests for Mendel's experiments

- In Mendel's pea plant experiments,
  yellow seeds (Y) are dominant and green (y) recessive;
  round seeds (R) are dominant and wrinkled (r) are recessive.

- Consider the phenotypes of the offspring in a "dihybrid cross" $YyRr \times YyRr$:

Type	Expected fraction	Observed number
yellow & round	9/16	315
yellow & wrinkled	3/16	101
green & round	3/16	108
green & wrinkled	1/16	32
	Total:	$n = 556$

- Hypothesis test:

  H_0: $(p_1, p_2, p_3, p_4) = (\frac{9}{16}, \frac{3}{16}, \frac{3}{16}, \frac{1}{16})$

  H_1: At least one p_i disagrees

Does the data fit the expected distribution?

Type	Expected fraction	Observed number
yellow & round	9/16	315
yellow & wrinkled	3/16	101
green & round	3/16	108
green & wrinkled	1/16	32
	Total:	$n = 556$

- The observed number of "yellow & round" plants is $O = 315$. (Don't confuse the letter O with the number 0.)
- The expected number is
$$E = (9/16) \cdot 556 = 312.75.$$
- The goodness of fit test requires that we convert all the expected proportions into expected numbers.

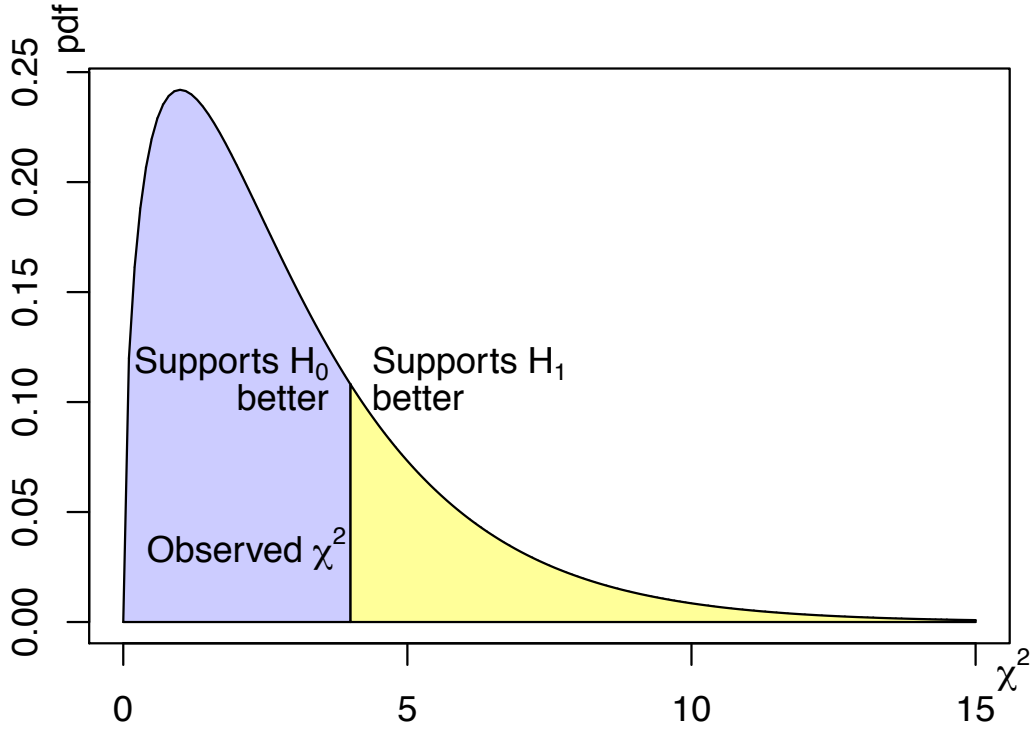
Goodness of fit test

Type	Observed number O	Expected number E	$O-E$	$(O-E)^2/E$
yellow & round	315	$(9/16)556 = 312.75$	2.25	0.0161871
yellow & wrinkled	101	$(3/16)556 = 104.25$	−3.25	0.1013189
green & round	108	$(3/16)556 = 104.25$	3.75	0.1348921
green & wrinkled	32	$(1/16)556 = 34.75$	−2.75	0.2176259
Total	556	556	0	0.4700240

- $k = 4$ categories give $k - 1 = 3$ degrees of freedom.
  (The O and E columns both total 556, so the $O - E$ column totals 0; thus, any 3 of the $(O - E)$'s dictate the fourth.)
- The test statistic is the total of the last column, $\chi^2_3 = 0.4700240$.
- The general formula is $\chi^2_{k-1} = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$.
- **Warning:** Technically, that formula only has an approximate chi-squared distribution. When $E \geqslant 5$ in all categories, the approximation is pretty good.

Goodness of fit test

- Smaller values of χ^2 indicate better agreement between the O and E values (so support H_0 better). Larger values support H_1 better. **It's a one-sided test.**



- The P-value is the probability, under H_0, of a test statistic that supports H_1 as well as or better than the observed value:

$$P = P(\chi^2_3 \geqslant 0.4700240) = .9254259$$

Matlab: `1-chi2cdf(.4700240,3)`

R: `1-pchisq(.4700240,3)`

Goodness of fit test

- $P = .9254259$ is not too extreme.
  It means that if H_0 is true and the experiment is repeated a lot, about 7.5% of the time, a χ^2_3 value supporting H_0 better (lower values of χ^2_3) will be obtained, and about 92.5% of the time, values supporting H_1 better (higher values of χ^2_3) will be obtained.

Ronald Fisher (1890–1962)

- He made important contributions to both statistics and genetics.
- Connection: he invented statistical methods while working on genetics problems.
- Our way of using the normal, Student t, χ^2, and F distributions in the same framework, plus ANOVA, is due to him.
- In genetics, he reconciled continuous variations (heights and weights) with Mendelian genetics (discrete traits), and developed much of population genetics.

Did Mendel fudge his data?

- For *independent experiments,* the values of χ^2 may be "pooled" by adding the χ^2 values and adding the degrees of freedom.
- Fisher pooled the data from Mendel's experiments and got $\chi^2 = 41.6056$ with 84 degrees of freedom.
- Assuming Mendel's laws are true, how often would we get χ^2_{84} supporting H_0/H_1 better than this?

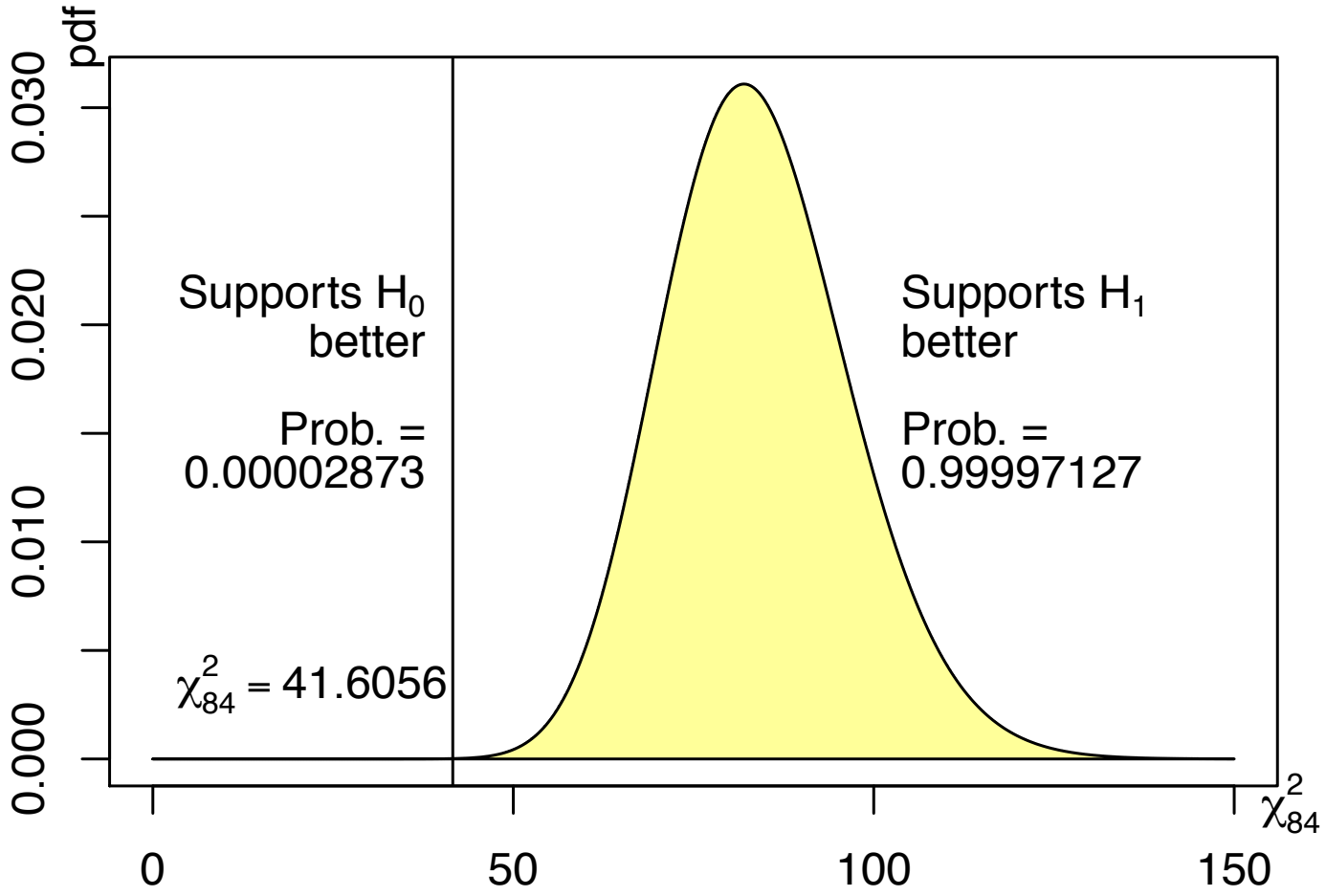
  Support H_0 better:

  $$P(\chi^2_{84} \leqslant 41.6056) = 0.00002873$$

  Support H_1 better:

  P-value $P = P(\chi^2_{84} \geqslant 41.6056) = 1 - 0.00002873 = .99997127.$
- So if Mendel's laws hold and 1 million researchers independently conducted the same experiments as Mendel, about 29 of them would get data with as little or even less variation than Mendel had.

Did Mendel fudge his data?



Did Mendel fudge his data?

Based on this and similar tests, Fisher believed that something was fishy with Mendel's data:

- The values are "too good" in the sense that they are too close to what was expected.
- At the same time, they are "bad" in the sense that there is too little random variation.
- Some people have accused Mendel of faking data.
- Others speculate that he only reported his best data.
- Other people defend Mendel by speculating on biological explanations for why his results would be better than expected.
- All pro and con arguments have later been rebutted by someone else.

Tests of independence ("contingency tables")

A study in 1899 examined 6800 German men to see if hair color and eye color are related.

Observed counts O:		Hair color				
		Brown	Black	Fair	Red	**Total**
Eye Color	Brown	438	288	115	16	857
	Gray/Green	1387	746	946	53	3132
	Blue	807	189	1768	47	2811
	Total	2632	1223	2829	116	6800

Hypothesis test (at $\alpha = 0.05$)

H_0: eye color and hair color are independent, vs.
H_1: eye color and hair color are correlated

Meaning of independence

For all eye colors x and all hair colors y:

$$P(\text{eye color}=x \text{ and hair color}=y) = P(\text{eye color}=x) \cdot P(\text{hair color}=y)$$

Computing E table

Hypothesis test (at $\alpha = 0.05$)

H_0: eye color and hair color are independent, vs.
H_1: eye color and hair color are correlated

- The fraction of people with red hair is $116/6800$.
- The fraction with blue eyes is $2811/6800$.
- Use these as point estimates: $P(\text{hair color=red}) \approx 116/6800$ and $P(\text{eye color=blue}) \approx 2811/6800$.
- Under the null hypothesis, the fraction with red hair and blue eyes would be $\approx (116 \cdot 2811)/6800^2$.
- The expected number of people with red hair and blue eyes is $6800(116 \cdot 2811)/6800^2 = (116 \cdot 2811)/6800 = 47.95$.
  (Row total times column total divided by grand total.)
- Compute E this way for all combinations of hair and eye color.
  As long as $E \geqslant 5$ in every cell (here it is) and the data is normally distributed (an assumption), the χ^2 test is valid.

Computing E and $O - E$ tables

Expected counts E:		Hair color			
		Brown	Black	Fair	Red
Eye Color	Brown	331.71	154.13	356.54	14.62
	Gray/Green	1212.27	563.30	1303.00	53.43
	Blue	1088.02	505.57	1169.46	47.95

In each position, compute $O - E$.
For red hair and blue eyes, this is $O - E = 47 - 47.95 = -.95$:

$O - E$:		Hair color			
		Brown	Black	Fair	Red
Eye Color	Brown	106.29	133.87	-241.54	1.38
	Gray/Green	174.73	182.70	-357.00	-0.43
	Blue	-281.02	-316.57	598.54	-0.95

Note all the row and column sums in the $O - E$ table are 0, so if we hid the last row and column, we could deduce what they are. Thus, this 3×4 table has $(3-1)(4-1) = 6$ degrees of freedom.

Computing test statistic χ^2

Compute $(O-E)^2/E$ in each position.
For red hair and blue eyes, this is $(-.95)^2/47.95 = 0.0189$.
(You could go directly to this computation after the E computation, without doing $O-E$ first.)

$(O-E)^2/E$:		Hair color			
		Brown	Black	Fair	Red
Eye	Brown	34.0590	116.2632	163.6301	0.1304
Color	Gray/Green	25.1852	59.2571	97.8139	0.0034
	Blue	72.5845	198.2220	306.3398	0.0189

Add all twelve of these to get

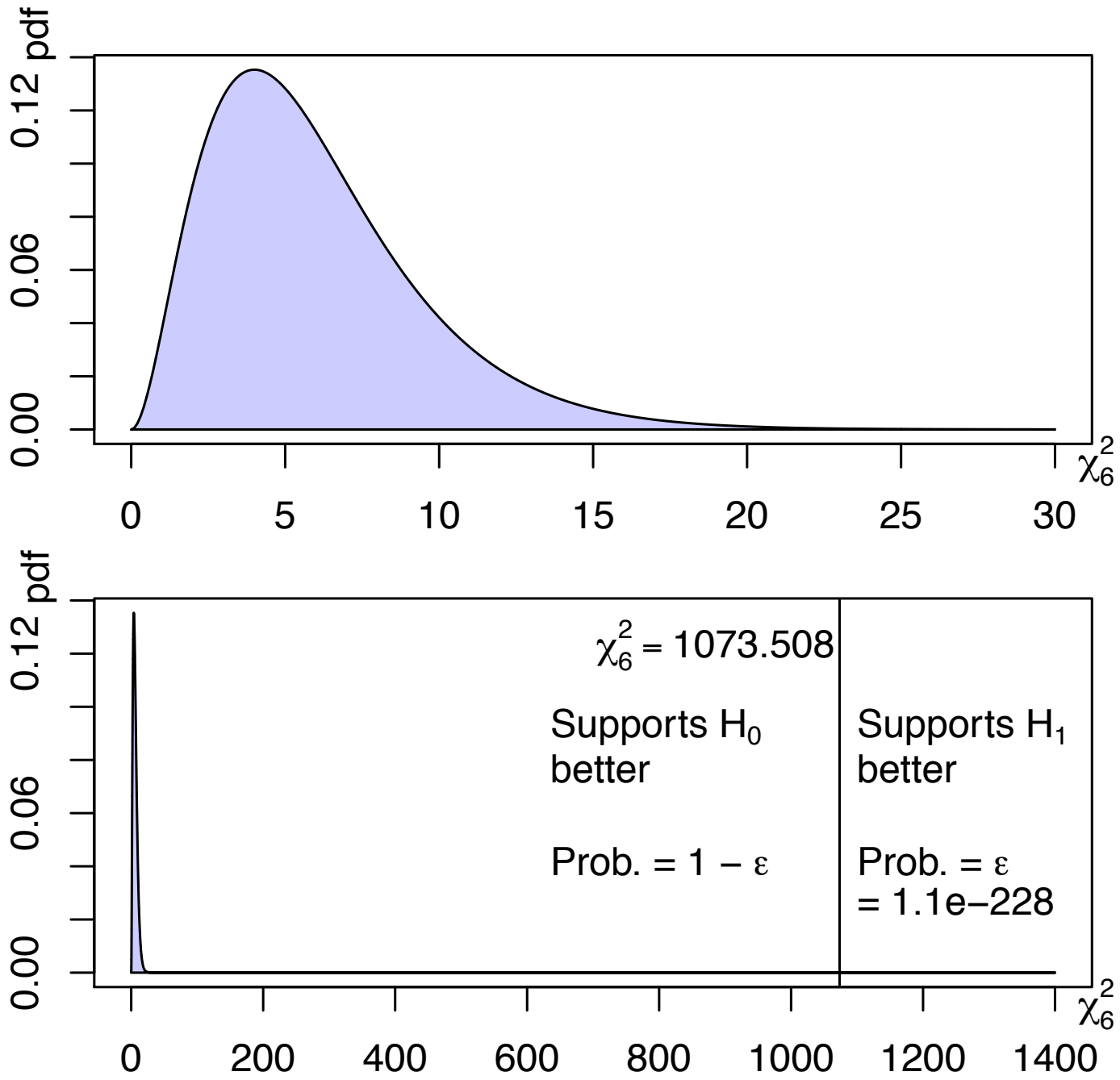
$$\chi^2 = 34.0590 + \cdots + 0.0189 = 1073.5076$$

There are 6 degrees of freedom, so $\chi^2_6 = 1073.5076$.

Performing the test of independence

- χ^2 would be 0 if the traits were truly independent.
  Smaller values support H_0 better (traits independent).
  Larger values support H_1 better (traits correlated).
  It's a one-sided test.
- At the 0.05 level of significance, we reject H_0 if
  $$\chi^2_6 \geqslant \chi^2_{0.95,6} = 12.5916$$
  Indeed, $1073.5076 > 12.5916$ so we reject H_0 and conclude that hair color and eye color are linked in this data.
- This doesn't prove that a particular hair color causes one to have a particular eye color, or vice-versa; it just says there's a correlation in this data.
- **Using P-values:** $P = P(\chi^2_6 \geqslant 1073.5076) \approx 1.1 \cdot 10^{-228}$
  so $P \leqslant \alpha = 0.05$ and we reject H_0.
  Matlab: can't compute this (gives $P = 0$).
  R: `pchisq(1073.5076,6,lower.tail=FALSE)`

Performing the test of independence



Mendel's pea plants revisited: Are loci Y and R linked?

We will use the same data as in the goodness-of-fit test but for a different purpose. Consider the phenotypes of the offspring in a "dihybrid cross" $YyRr \times YyRr$:

Observed counts O:		Seed Shape		
		Round (R)	Wrinkled (r)	**Total**
Seed Color	Yellow (Y)	315	101	416
	Green (y)	108	32	140
	Total	423	133	556

Hypothesis test (at $\alpha = 0.05$)

H_0: Seed color and seed shape are independent, vs.
H_1: Seed color and seed shape are correlated

Mendel's pea plants revisited: Are loci Y and R linked?

	Seed Color	Seed Shape		
		Round (R)	Wrinkled (r)	**Total**
O: (Observed #)	Yellow (Y)	315	101	416
	Green (y)	108	32	140
	Total	423	133	556
E: (Expected #)	Yellow (Y)	316.4892	99.5108	416
	Green (y)	106.5108	33.4892	140
	Total	423	133	556
$O-E$: (Deviation)	Yellow (Y)	-1.4892	1.4982	0
	Green (y)	1.4892	-1.4892	0
	Total	0	0	0
$(O-E)^2/E$: (χ^2 contrib.)	Yellow (Y)	0.0070	0.0223	0.0293
	Green (y)	0.0208	0.0662	0.0870
	Total	0.0278	0.0885	0.1163

Mendel's pea plants revisited: Are loci Y and R linked?

Using χ^2 as the test statistic:

$$df = (2-1)(2-1) = 1$$

$$\chi^2_1 = .0070 + .0223 + .0208 + .0662 = 0.1163$$

cutoff: $\chi^2_{0.95,1} =$ `chi2inv(.95,1)` $=$ `qchisq(.95,1)` $= 3.8415$

$0.1163 < 3.8415$ so it's not significant

Using P-values:

$$P = P(\chi^2_1 > 0.1163) =$$ `1-chi2cdf(0.1163,1)` $= 0.7331$

`1-pchisq(0.1163,1)`

$P > 0.05$ so Accept H_0 (genes not linked)

Comparison of the two tests

- At fertilization, if genes R and Y are not linked, then in an $RrYy \times RrYy$ cross, the expected proportions are
$$RY:Ry:rY:ry = 1:1:1:1.$$
If linked, it would be different.

- Some genotypes may not survive to the points at which the phenotype counts are made; e.g., hypothetically, 40% of individuals with Rr might not be born, might die before reproducing (affecting multigenerational experiments), etc.

  This would change the ratio of
$$RR:Rr:rr \quad \text{from} \quad 1:2:1 \quad \text{to} \quad 1:1.2:1 = 5:6:5,$$
and round:wrinkled from $3:1$ to $2.2:1 = 11:5$.

Comparison of the two tests

- The goodness-of-fit test assumed all genotypes are equally viable.

  Whether the genes are linked or not should be a separate matter.

  If you know the yellow:green and round:wrinkled viability ratios, you can use the goodness-of-fit test on 4 phenotypes with 3 degrees of freedom by adjusting the proportions.

- If you don't know these viability ratios, you can estimate the ratios from data via contingency tables, at the cost of dropping to 1 degree of freedom.