Lecture 7: Discrete data problems I

Pratheepa Jeganathan

04/17/2019

Recall

- One sample sign test, Wilcoxon signed rank test, large-sample approximation, median, Hodges-Lehman estimator, distribution free confidence interval.
- Jackknife for bias and standard error of an estimator.
- Bootstrap samples, bootstrap replicates.
- Bootstrap standard error of an estimator.
- Bootstrap percentile confidence interval.
- Hypothesis testing with the bootstrap (one-sample problem.)
- Assessing the error in bootstrap estimates.
- Example: inference on ratio of heart attack rates in aspirin-intake group to placebo group.
- The exhaustive bootstrap distribution

Discrete random variable with two categories

A binomial test

- Let X_i be a (Bernoulli) random variable (two categories - success/failure) with success probability p.
  - $\mathbb{E}(X_i) = p$ and $\mathbb{V}(X_i) = p(1-p)$.
- Statistical problems:
  - Hypothesis testing on p.
  - Confidence interval for p.
  - Estimator for p.
- Let $B = \sum_{i=1}^{n} X_i$ be the total number of success.
- $B \sim \text{Binomial}(n, p)$.

A binomial test

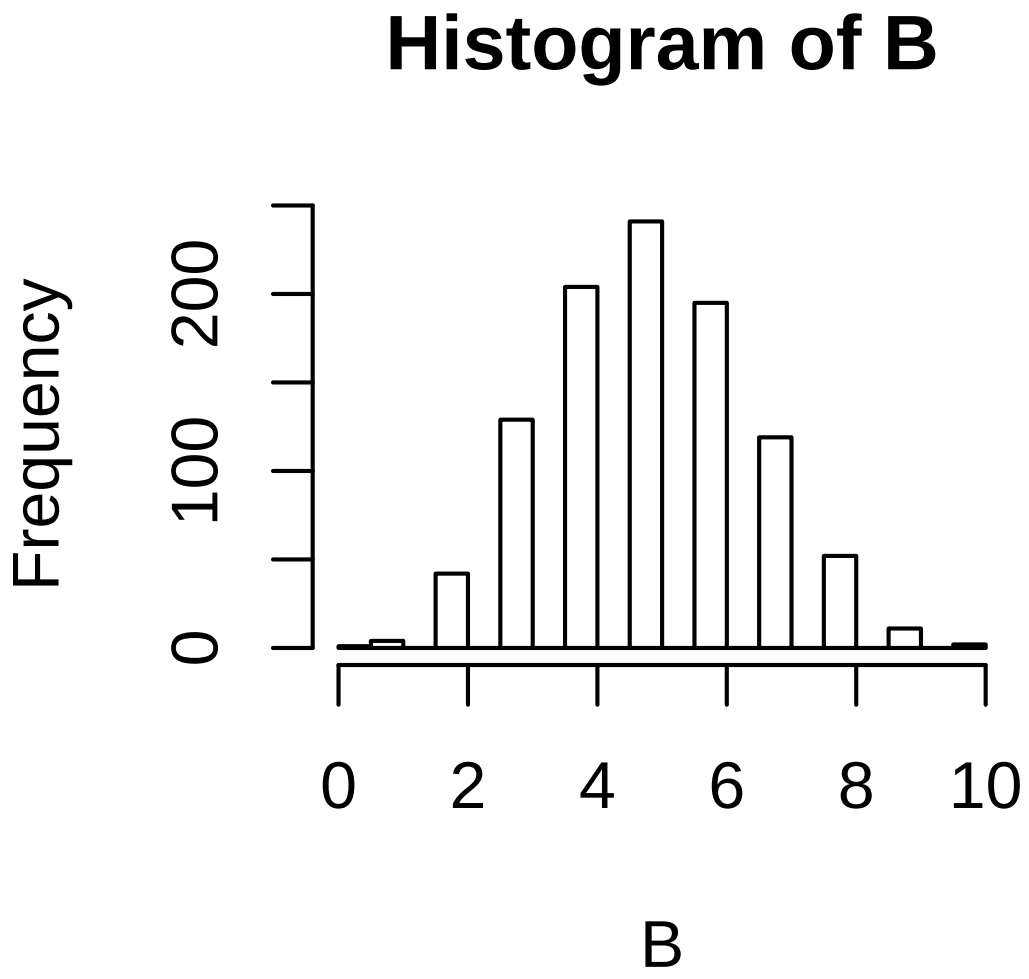
- Hypothesis test: $\mathrm{H}_0 : p = p_0$ versus $\mathrm{H}_A : p \neq p_0$.
  - Test statistic: $B = \sum_{i=1}^{n} X_i \sim \text{Binomial}\,(n, p_0)$.
  - Rejection regions
    - $\mathrm{H}_A : p > p_0$, Reject H_0 if $B \geq b_{\alpha;n,p_0}$.
    - $\mathrm{H}_A : p < p_0$, Reject H_0 if $B \leq c_{\alpha;n,p_0}$.
    - $\mathrm{H}_A : p \neq p_0$, Reject H_0 if $B \geq b_{\alpha_1;n,p_0}$ or $B \leq c_{\alpha_2;n,p_0}$, where $\alpha_1 + \alpha_2 = \alpha$.
- Due to discreteness of B, we cannot do test for all α values.

A binomial test (large-sample approximation)

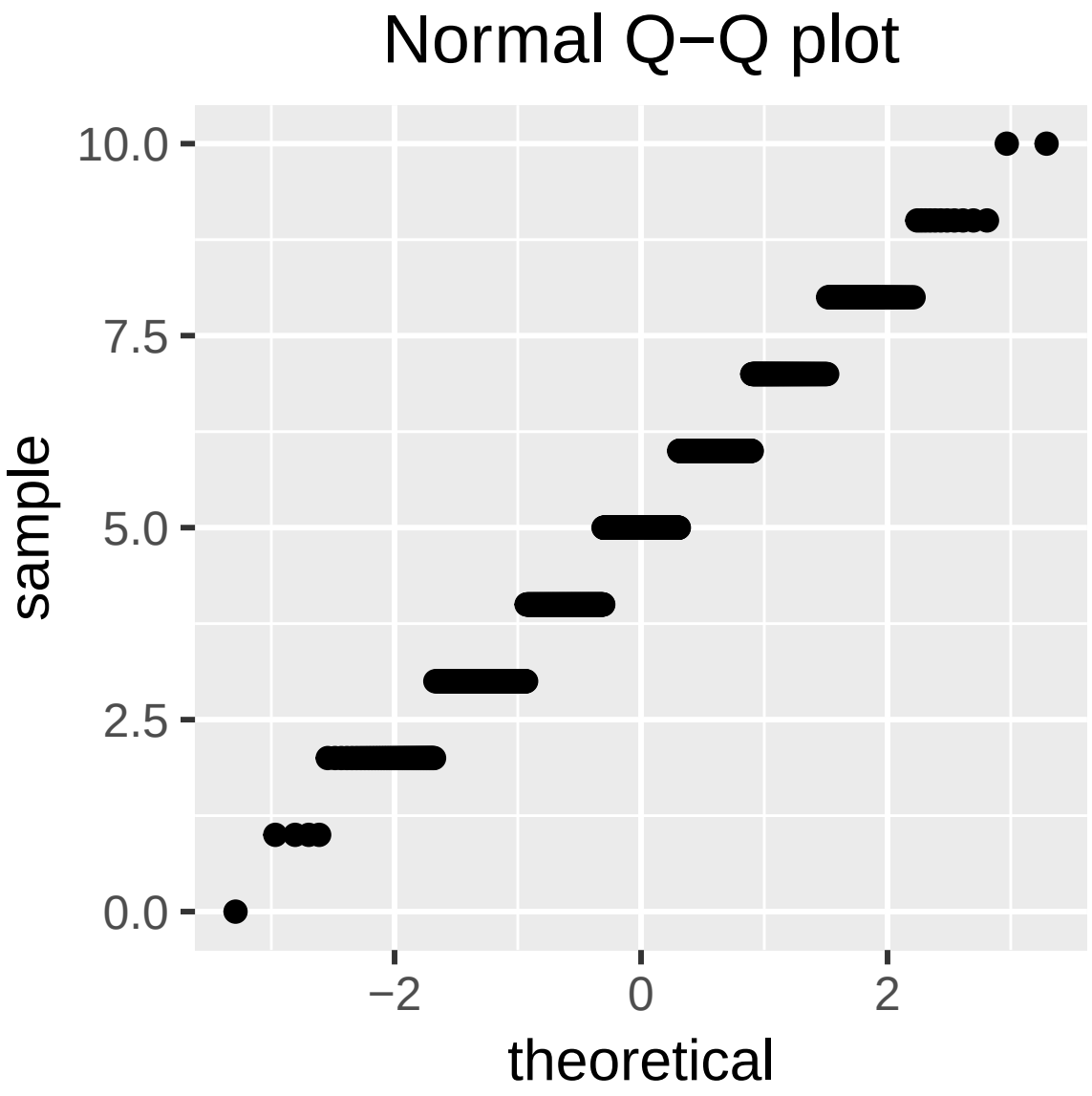
```
n = 10; p0 = 1/2; nsim = 1000
B = rbinom(nsim, size = n, prob = p0)
```

A binomial test (large-sample approximation)

```
hist(B, breaks = 30)
```



A binomial test (large-sample approximation)

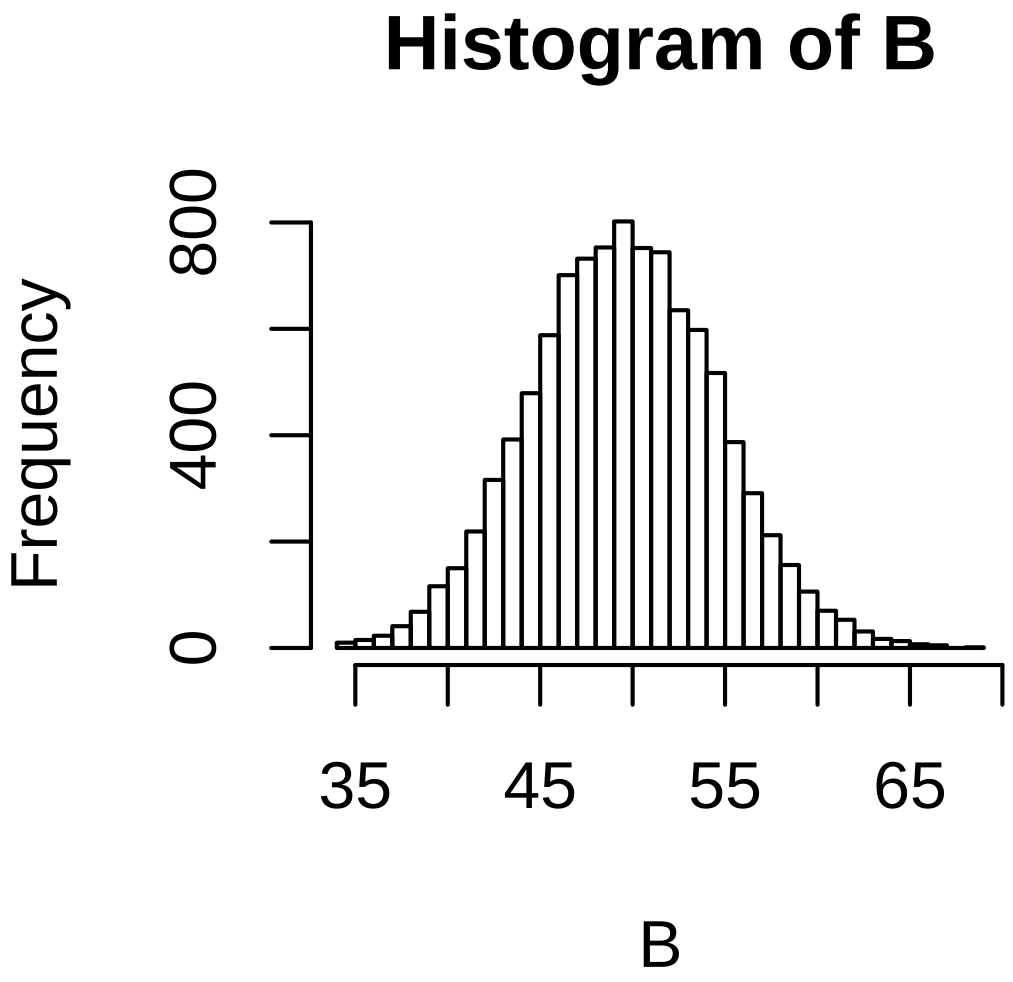


A binomial test (large-sample approximation)

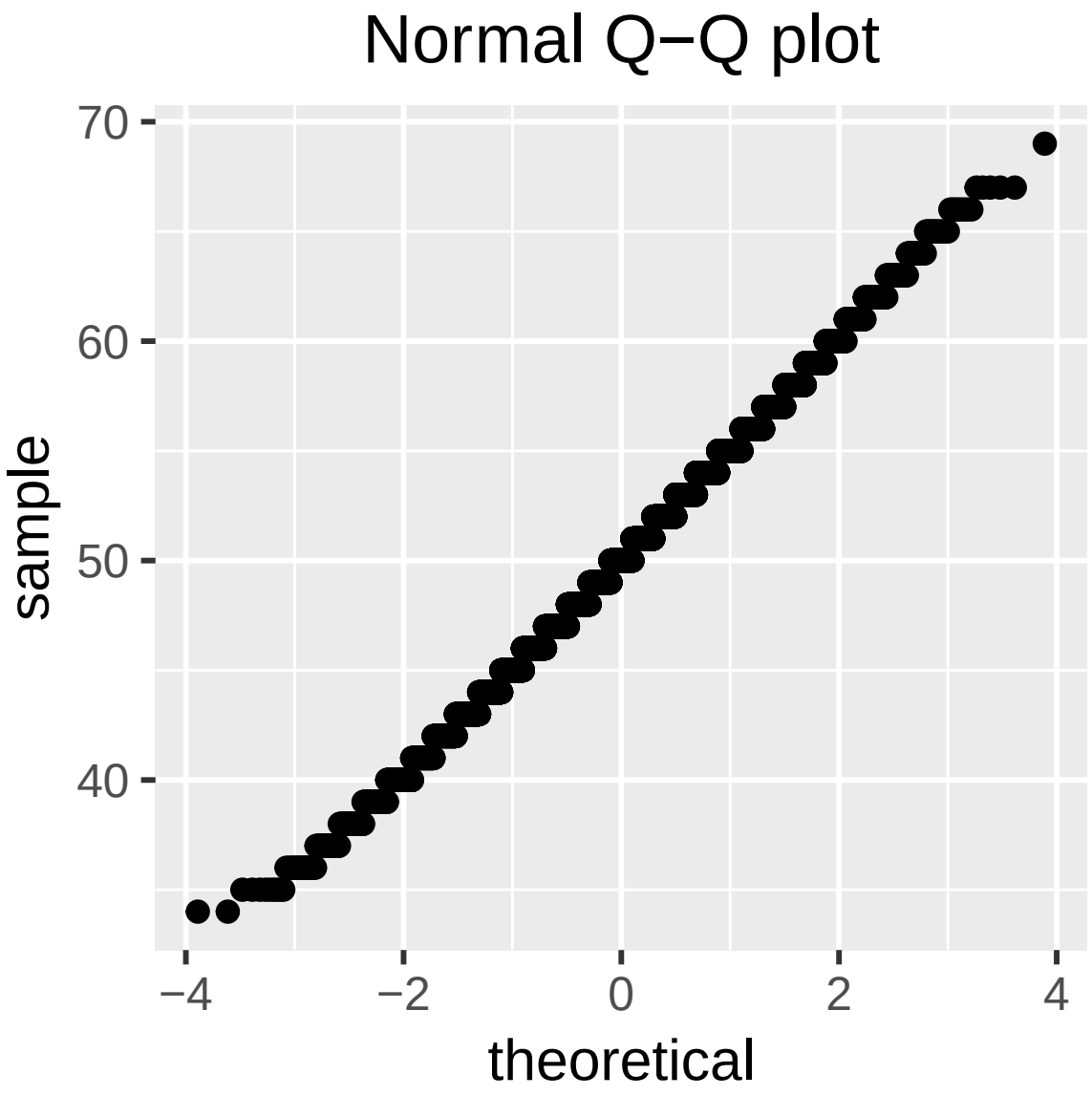
```
n = 100; p0 = 1/2; nsim = 10000
B = rbinom(nsim, size = n, prob = p0)
```

A binomial test (large-sample approximation)

```
hist(B, breaks = 30)
```



A binomial test (large-sample approximation)



A binomial test (large-sample approximation)

- When $\mathrm{H}_0 : p = p_0$ is true
  - $\mathbb{E}(B) = np_0$ and
  - $\mathbb{V}(B) = np_0(1 - p_0)$
  - The standardized version of B is $Z = \dfrac{B - np_0}{(np_0(1 - p_0))^{1/2}}$ and
  - $Z \sim N(0, 1)$.
  - $Z^2 \sim \chi^2_{df=1}$, where df is degrees of freedom.

A binomial test (large-sample approximation)

- When $n \to \infty$, while $p = p_0$ is fixed under H_0, test statistic $Z \sim N(0, 1)$
- Rejection regions
  - $\mathrm{H}_A : p > p_0$, Reject H_0 if $Z \geq z_\alpha$.
  - $\mathrm{H}_A : p < p_0$, Reject H_0 if $Z \leq -z_\alpha$.
  - $\mathrm{H}_A : p \neq p_0$, Reject H_0 if $Z \geq z_{\alpha/2}$ or $Z \leq -z_{\alpha/2}$.

Example: Sensory Difference Tests

- Experiment:
  - To each of n panelists, three test samples are presented in a randomized order.
  - Two of the samples are known to be identical; the third is different. The panelist is then supposed to select the odd sample.
  - Assume panelists are homogeneous trained judges, so experiment has n Bernoulli trials.
  - Let p success probability corresponds to a correct identification of the odd sample.
- Test the hypothesis that there is a basis for discrimination (i.e. $\mathrm{H}_A : p > \frac{1}{3}$).

Example: Sensory Difference Tests (use rejection region)

- Data:
  - Out of 50 trials, there were 25 correct selections and 25 incorrect selections.
- $H_0 : p = \frac{1}{3}$ versus $H_A : p > \frac{1}{3}$.
- Test statistic $Z = \frac{B - 50\left(\frac{1}{3}\right)}{\left(50\left(\frac{1}{3}\right)\left(\frac{2}{3}\right)\right)^{1/2}}$. (large-sample approximation)
- Significance level: $\alpha = .05$.

```
qnorm((1-.05), mean = 0, sd =1)
```

```
## [1] 1.644854
```

- Rejection region: $Z \geq z_{.05} = 1.645$.

Example: Sensory Difference Tests

- Observed test statistic $Z_o = \dfrac{25 - 50\left(\frac{1}{3}\right)}{\left(50\left(\frac{1}{3}\right)\left(\frac{2}{3}\right)\right)^{1/2}} = 2.5.$

```
Z.obs = (25 - 50*1/3)/(sqrt(50*1/3*2/3)); Z.obs
```

```
## [1] 2.5
```

- The large sample approximation value $Z_o = 2.5 > 1.645$ and thus we reject $H_0 : p = \frac{1}{3}$ in favor of $p > \frac{1}{3}$ at the approximate $\alpha = .05$ level. Thus there is evidence of a basis for discrimination in the sensory test.

Example: Sensory Difference Tests (use p-value)

- P -value corresponding to observed test statistic value $Z_o = 2.5$ is $P(Z \geq 2.5)$

```
1 - pnorm(2.5)
```

```
## [1] 0.006209665
```

- Thus, the smallest significance level at which we reject H_0 in favor of $p > \frac{1}{3}$ using the large-sample approximation is .0062.
- The exact P-value in this case is $P(B \geq 25) = 1 - P(B \leq 24)$

```
1-pbinom(24,50,1/3)
```

```
## [1] 0.01082668
```

Calculating Power

- Suppose we have $n = 50$ and we decide to employ the approximate $\alpha = .05$ level test of $H_0 : p = \frac{1}{3}$ versus $H_A : p > \frac{1}{3}$.
- We found that test reject H_0 is $Z \geq 1.645$.
- What is the power of this test if in fact $p = .6$?
  - Power is the probability of rejecting H_0 when H_A is true.

Calculating Power

- Now $p = .6$, $Z = \dfrac{B - 50\left(\frac{1}{3}\right)}{\left(50\left(\frac{1}{3}\right)\left(\frac{2}{3}\right)\right)^{1/2}}$ is no longer standard normal.
- We have $Z^* = \dfrac{B - 50\,(.6)}{(50\,(.6)\,(.4))^{1/2}} \sim \mathrm{N}\,(0,1)$.

$$\begin{aligned}
\text{Power} &= P\,(Z \geq 1.645 | p = .6) \\
&= P_{p=.6}\left(\frac{B - 50\left(\frac{1}{3}\right)}{\left(50\left(\frac{1}{3}\right)\left(\frac{2}{3}\right)\right)^{1/2}} \geq 1.645\right) \\
&= P_{p=.6}\left(B \geq 1.645\left(50\left(\frac{1}{3}\right)\left(\frac{2}{3}\right)\right)^{1/2} + 50\left(\frac{1}{3}\right)\right) \\
&= P_{p=.6}\left(\frac{B - 50\,(.6)}{(50\,(.6)\,(.4))^{1/2}} \geq \frac{1.645\left(50\left(\frac{1}{3}\right)\left(\frac{2}{3}\right)\right)^{1/2} + 50\left(\frac{1}{3}\right) + 50\,(.6)}{(50\,(.6)\,(.4))^{1/2}}\right) \\
&= P\left(Z^* \geq -2.27\right) = .9884.
\end{aligned}$$

An estimator for probability of success

- The estimator of the probability of success p, associated with the statistic B, is $\hat{p} = \frac{B}{n}$.
- Standard error of $\hat{p}$ is $\sqrt{\frac{p(1-p)}{n}}$ and estimate is $\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$.

Confidence interval for probability of success

- The large-sample $(1-\alpha)100\%$ confidence interval for p is $\left(\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)$, where $z_{\alpha/2}$ is the upper $\alpha/2$ quantile of standard normal distribution.

```
library(binom)
binom.confint(x=25, n=50, conf.level=.95, methods = "asympt
```

```
##       method  x  n mean     lower     upper
## 1 asymptotic 25 50  0.5 0.3614096 0.6385904
```

Discrete random variable with more than two categories

Pearson's Chi-Squared Goodness-of-Fit Test

- χ^2 test for specified multinomial probabilities.
- Let n experiments with frequencies $X_1, \cdots, X_k$ corresponding to the k categories.
- Test the hypothesis that the multinomial probabilities $p_1, \cdots, p_k$ are equal to specified or known values $p_1^0, \cdots, p_k^0$.
- $H_0 : p_1 = p_1^0, \cdots, p_k = p_k^0$ versus $H_A : p_i \neq p_i^0$ for at least one value i.
- Pearson's chi-squared statistic
$$\chi^2 = \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}}.$$
- Pearson's chi-squared statistic, in notation,
$$\chi^2 = \sum_{i=1}^{k} \frac{(X_i - np_i^0)^2}{np_i^0}.$$
- A large-sample approximation (when $np_i^0 \geq 5$ for each i)
  - As $n \to \infty$, χ^2 is that of a chi-squared distribution with $k-1$ degrees of freedom
- Rejection region: reject H_0 if $\chi^2 \geq \chi^2_{\alpha,k-1}$.

Example (Outcomes of Pea Plant Experiments)

- Gregor Mendel' s famous genetics experiments on pea plants.
- Experiment:
  - Cross-pollinated purebred plants with specific traits and observed and recorded the results over many generations.
  - seed shape (round or angular), cotyledon (part of the embryo within the seed) color (yellow or green), seed coat color (colored or white), pod shape (inflated or constricted), pod color (green or yellow), flower position (axial or terminal), stem length (long or short).

Example (Outcomes of Pea Plant Experiments)

- Contingency table

```
df = data.frame(Dominant = c(5474,6022,705,882,428,651,787),
  Recessive = c(1850,2001,224,299,152,207,277))
rownames(df) = c("Seed_shape", "Cotyledon_color",
  "Seed_coat_color", "Pod_shape", "Pod_color",
  "Flower_position", "Stem_length"); df
```

```
##                   Dominant Recessive
## Seed_shape            5474      1850
## Cotyledon_color       6022      2001
## Seed_coat_color        705       224
## Pod_shape              882       299
## Pod_color              428       152
## Flower_position        651       207
## Stem_length            787       277
```

Example (Outcomes of Pea Plant Experiments)

- Goodness-of-fit test
- $H_0 : p_{1d} = p_{2d} = \cdots = p_{7d} = \frac{3}{4}$ versus $H_A : p_{id} \neq \frac{3}{4}$ for at least one i, p_{id} is the probability of the second offspring of cross-pollinated purebred plant have dominant characteristic.

```
library(dplyr); df = mutate(df,
  expected.ratio = rep("3:1",times = 7));
rownames(df) = c("Seed_shape", "Cotyledon_color", "Seed_coa
```

```
##                   Dominant Recessive expected.ratio
## Seed_shape            5474      1850            3:1
## Cotyledon_color       6022      2001            3:1
## Seed_coat_color        705       224            3:1
## Pod_shape              882       299            3:1
## Pod_color              428       152            3:1
## Flower_position        651       207            3:1
## Stem_length            787       277            3:1
```

Example (Outcomes of Pea Plant Experiments)

- Chi-squared statistic for each row.

```
chi.sq = apply(df[,1:2], 1, function(x){
  chisq.test(c(x[1],x[2]),
    p = c(.75, .25))$statistic
}); df = mutate(df, chi.sq = chi.sq); df
```

```
##   Dominant Recessive expected.ratio     chi.sq
## 1     5474      1850            3:1 0.26288003
## 2     6022      2001            3:1 0.01499855
## 3      705       224            3:1 0.39074273
## 4      882       299            3:1 0.06350550
## 5      428       152            3:1 0.45057471
## 6      651       207            3:1 0.34965035
## 7      787       277            3:1 0.60651629
```

Example (Outcomes of Pea Plant Experiments)

- Each row has a $\chi^2_{df=1}$ distribution with degrees of freedom (df) 1.
- Sum seven independent χ^2 random variables with each df 1 gives a $\chi^2_{df=7}$ with degrees of freedom (df) 7.

```
sum(df$chi.sq)
```

```
## [1] 2.138868
```

Where does the observed chi-squared value fall?

```
pchisq(sum(df$chi.sq), df = 7, lower.tail = TRUE)
```

```
## [1] 0.04824407
```

- The value 2.1389 falls in the lower tail of the distribution. Thus, we do not have enough evidence to reject H_0.

Example (Outcomes of Pea Plant Experiments)

- Mendel did many more experiments than the data we used.
- Fisher suspected that an overzealous assistant might have biased the data.
- Over time the works of Mendel and many others have led to acceptance of Mendel's genetic theories.
- Read more about Mendel's genetics here

References for this lecture

HWC Chapter 2

HWC Chapter 2, page 29, comment 26 (Pearson's Chi-Squared Goodness-of-Fit Test)