Inference Overview

Recall that random variables are characterized by their distributions, which often involve unknown parameters.

- By inference, we mean to use **data** to
  1. **estimate** parameters of interest
  2. find **confidence intervals** for parameters
  3. **test hypotheses** about parameters

- To estimate parameters, we may use method of moments (MME) or maximum likelihood estimator (MLE) etc. Some simple estimators can be intuitively derived, eg. use sample mean to estimate population mean.

- To achieve 2 and 3, we need to know the **distribution** of the test statistic or parameter estimate

- In some cases, we are able to get the **exact** distribution, such as Fisher's Exact Test.

- In other cases, we must use large sample theory to get an **asymptotic**, or approximate, distribution.

We can often use large sample theory to show that some parameter estimates and test statistics are **normally distributed**. In these situations, we really only need to know the MEAN and VARIANCE of the random variables (since the mean and variance completely specify the Normal distribution).

The Normal Distribution

The normal distribution is an example of a **probability distribution** for a continuous random variable.

- It is specified by its **density:**

$$f(y) \;=\; \frac{1}{\sqrt{2\,\pi\,\sigma^2}}\, e^{\frac{-(y-\mu)^2}{2\,\sigma^2}}$$

- There is a whole family of normal distributions, denoted by $N(\mu, \sigma^2)$, specified by values of the parameters μ and σ.

  μ = mean of the population distribution

  σ = standard deviation of the population distribution

- The **Standard Normal Distribution** is defined by $\mu = 0$ and $\sigma = 1$. The density can thus be simplified to:

$$f(y) \;=\; \frac{1}{\sqrt{2\,\pi}}\, e^{\frac{-y^2}{2}}$$

Normal distribution review:

- To calculate the probability that a $N(0,1)$ r.v. Z falls in the interval from a to b, we could use calculus:

$$Pr(a \leq Z \leq b) \;=\; \int_a^b \frac{1}{\sqrt{2\,\pi}}\, e^{\frac{-y^2}{2}} dy$$

- The probabilities $Pr(Z \leq z)$ for various values of z are also given by statistical tables or software [R: pnorm()]. Then we can calculate the above probability such as:

$$Pr(a \leq Z \leq b) \;=\; Pr(Z \leq b) - Pr(Z \leq a)$$

- For normal r.v.'s with mean μ and standard deviation σ, traditionally we use the following **standardization** to calculate probabilities:

$$\begin{aligned} Pr(a \leq Y \leq b) &= Pr\left(\frac{a-\mu}{\sigma} \leq \frac{Y-\mu}{\sigma} \leq \frac{b-\mu}{\sigma}\right) \\ &= Pr(a^* \leq Z \leq b^*) \\ &= Pr(Z \leq b^*) - Pr(Z \leq a^*) \end{aligned}$$

  where Z is $N(0,1)$, $a^* = (a-\mu)/\sigma$, and $b^* = (b-\mu)/\sigma$.

  Note: you don't need this step using software.

- All normal distributions have 95% of their area between $(\mu - 1.96\sigma)$ and $(\mu + 1.96\sigma)$.

Large Sample Theory

Central Limit Theorem (CLT) :

Let Y be the sum of n independent, identically distributed (i.i.d.) random variables $Y_1, Y_2, ..., Y_n$:

$$Y = \sum_{i=1}^{n} Y_i$$

Then, for large n,

$$Z = \left(\frac{Y - E(Y)}{\sqrt{\text{Var}(Y)}} \right) \overset{approx}{\sim} N(0, 1),$$

- There are certain "regularity" conditions that must be satisfied, such as
$$0 < \text{Var}(Y_i) < \infty.$$

- Most of the statistical tests we perform are based on the Central Limit Theorem.

Another form of the CLT:

Let $\overline{Y}$ be the sample mean,

$$\overline{Y} = \frac{\Sigma_{i=1}^{n} Y_i}{n},$$

of n i.i.d. random variables $Y_1, Y_2, ..., Y_n$ with

$$E(Y_i) = \mu \quad \text{and} \quad \text{Var}(Y_i) = \sigma^2$$

Then, for large n,

$$Z = \left(\frac{\overline{Y} - \mu}{\sigma/\sqrt{n}} \right) \dot{\sim} N(0, 1).$$

Here, we have stated the Central Limit Theorem in terms of the sample mean, instead of the sum.

Example: Binomial Data $Y \sim Bin(n, p)$

- A usual estimator for p (why):

$$\hat{p} = \overline{Y} = \frac{Y}{n} = \frac{\Sigma_{i=1}^{n} Y_i}{n},$$

  where the Y_i are **i.i.d.** Bernoulli random variables.

- We know that the "exact" distribution of Y is

$$Y = n\hat{p} \sim Bin(n, p)$$

  What is $P(Y = k) =$?

- Note that $\hat{p}$ is just the sample mean of the Bernoulli r.v.'s

- To apply CLT, we have n i.i.d Bernoulli r.v.'s with

$$E(Y_i) = \mu = p$$

  and

$$\text{Var}(Y_i) = \sigma^2 = p(1 - p)$$

- Substituting $\overline{Y} = \hat{p}$, $\mu = p$ and $\sigma^2 = p(1-p)$ in the CLT, we get:

$$\begin{aligned} Z &= \frac{\overline{Y} - \mu}{\sqrt{\sigma^2/n}} \\ &= \frac{\hat{p} - p}{\sqrt{p(1-p)/n}} \\ &\dot{\sim} N(0,1) \end{aligned}$$

  when n is large.

- In other words, for large n we can say:

$$\hat{p} \dot{\sim} N\left(p, \frac{p(1-p)}{n}\right)$$

Notes:

- For large n, confidence intervals and test statistics based on the exact distribution,

$$n\hat{p} \sim Bin(n, p)$$

can be cumbersome (computationally intensive)

- Furthermore, they are almost identical to those based on

$$\hat{p} \dot{\sim} N\left(p, \frac{p(1-p)}{n}\right),$$

so the latter is usually used for large n.

- If n is large, and neither p nor $(1-p)$ is close to 0, then the normal approximation works well.

- The closer p or $(1-p)$ is to 0, the worse the normal approximation (you need larger n for the normal approximation to be OK).

- The typical assumption for the normal approximation to be good is

$$0 < p - 3\sqrt{\frac{p(1-p)}{n}} < p + 3\sqrt{\frac{p(1-p)}{n}} < 1.$$

Transformations of r.v. (Delta Method)

When we looked at the CLT, we stated it in terms of both the sum and the sample mean.

Now suppose we want the approximate distribution of the estimated 'logit': $\text{logit}(\hat{p}) = \log\{\hat{p}/(1-\hat{p})\}$. (why)

The Delta Method: Suppose we have an asymptotically normal r.v. Y:

$$Y \dot{\sim} N(\mu, \sigma^2),$$

then

$$g(Y) \dot{\sim} N\left(g(\mu), [g'(\mu)]^2\, \sigma^2\right)$$

- Two regularity conditions are:

  $g(y)$ is differentiable

  $g'(\mu) \neq 0$

Example: The "Logit"

- Suppose $Y \sim B(n, p)$.

- Based on the CLT, we have the following large sample distribution:

$$\hat{p} \sim N\left(p, \frac{p(1-p)}{n}\right)$$

- Suppose we want to find the approximate distribution of the estimated 'logit':

$$g(\hat{p}) = \text{logit}(\hat{p}) = \log\left(\frac{\hat{p}}{1-\hat{p}}\right) = \log(\hat{p}) - \log(1-\hat{p})$$

- Using the Delta Method,

$$g(\hat{p}) \dot{\sim} N\left(g(p),\, [g'(p)]^2\, \frac{p(1-p)}{n}\right)$$

or, equivalently,

$$\text{logit}(\hat{p}) \sim N\left(\text{logit}(p),\ \frac{1}{np(1-p)}\right)$$

Confidence Intervals

- An estimate such as $\hat{p}$ comes with variability or uncertainly [see plot: OI biostat page 176], and we refer to it as a 'point estimate'.

- A confidence interval (CI) for a parameter θ (our general notation) is a **random interval**, computed from the data (hence random), that contains θ with pre-specified probability, eg. 95%.

  Interpretation of CI:
  This relies on the abstract construct of *repeated sampling*; i.e. if we take repeated random samples of the data (say $X_1, ..., X_n$) from the same population, and form a CI from each sample, then about 95% of them should contain θ. [see plot: OI biostat page 181]

- More generally, instead of 95%, we may consider a $100(1-\alpha)$% CI, and $(1-\alpha)$ is sometimes called the coverage probability.

- A CI gives a plausible *range of values* for θ with a *margin of error*.

CATEGORICAL OUTCOMES
Analysis of 2×2 contingency tables

Categorical vs. continuous variables

For many analysis purposes, for example:

- descriptive statistics ('Table 1'):
  - continuous variables are summarized by mean (SD)
  - categorical variables are summarized by count (%)

- regression models:
  - continuous outcomes are often modeled using linear regression;
  - categorical outcomes are often modeled using some type of logistic regression.

Cold incidence among French Skiers:

		OUTCOME			
		COLD	NO COLD	Total	
TREATMENT	VITAMIN C	17	122	139	<-- (fixed by design)
	NO VITAMIN C	31	109	140	<-- (fixed by design)
	Total	48	231	279	

- Number on each treatment fixed by design.
- Usually the design for experimental studies/clinical trials
- Individuals are followed to assess response

Questions:

1. what is the research question of interest?
   what is the outcome?
2. what are the distributions involved?
   (what are the random variables?)
3. what would be your hypotheses?
   (what are the parameters?)

In general, we can form the following 2×2 table:

		Outcome 1	Outcome 2	
TREATMENT	1	Y_1	$n_1 - Y_1$	n_1
	2	Y_2	$n_2 - Y_2$	n_2

- Individuals are given (or sometimes randomized to) treatment 1 or treatment 2

- The measured outcome is success or failure.

Facts about the distribution of outcomes

- n_1 and n_2 are fixed by design

- Y_1 and Y_2 are *independent* with distributions:

$$Y_1 \sim B(n_1, p_1)$$

$$Y_2 \sim B(n_2, p_2)$$

- We want to estimate p_1, p_2 and compare them.

- The *likelihood*, i.e. probability of observed data, is the product of 2 independent binomials:

$$\begin{aligned} L(p_1, p_2) &= P(Y_1 = y_1|p_1)\ P(Y_2 = y_2|p_2) \\ &= \left(\begin{array}{c} n_1 \\ y_1 \end{array} \right) p_1^{y_1}(1-p_1)^{n_1-y_1} \left(\begin{array}{c} n_2 \\ y_2 \end{array} \right) p_2^{y_2}(1-p_2)^{n_2-y_2} \end{aligned}$$

- Research question(s) of interest:
  1. Does treatment affect outcome? (causal)
  2. Are treatment and outcome associated? (observational)
  3. Is the success probability the same on both treatments?

- These are often assessed via a null hypothesis:

$$\mathrm{H}_0\text{: } p_1 = p_2 = p$$

  and an alternative hypothesis

$$\mathrm{H}_A\text{: } p_1 \neq p_2$$

- Two-sided alternatives are often considered more rigorous, because they are harder to reject (**why**).

- We are interested in
  1. describing treatment differences
  2. testing for a treatment effect.

Q: How can we quantify treatment differences?

Measuresof treatment differences

1. **Risk Difference**

$$\Delta = p_1 - p_2, \qquad -1 \leq \Delta \leq 1$$

2. **Relative Risk or Risk Ratio**

$$RR = \frac{p_1}{p_2}, \qquad 0 \leq RR \leq \infty$$

The log-relative risk is often used to get around the restriction that the relative risk must be positive:

$$\log RR = \log\left(\frac{p_1}{p_2}\right) = \log(p_1) - \log(p_2)$$

where

$$-\infty \leq \log RR \leq \infty.$$

3. **Odds Ratio or Relative Odds**

- The odds of success versus failure on treatment i is:

$$\frac{p_i}{(1-p_i)}, \qquad i = 1, 2$$

- The ratio of the odds for treatment 1 to treatment 2 is:

$$OR = \frac{p_1/(1-p_1)}{p_2/(1-p_2)} = \frac{p_1(1-p_2)}{p_2(1-p_1)}, \qquad 0 \leq OR \leq \infty,$$

- Again, the log-odds ratio is often used, to avoid the restriction that the odds ratio must be positive, i.e.,

$$\begin{aligned} \log OR &= \log\left(\frac{p_1/(1-p_1)}{p_2/(1-p_2)}\right) \\ &= \log\left(\frac{p_1}{1-p_1}\right) - \log\left(\frac{p_2}{1-p_2}\right) \\ &= \text{logit}(p_1) - \text{logit}(p_2) \end{aligned}$$

  where $-\infty \leq \log OR \leq \infty$

- Note that the $\log(OR)$ is the difference in logits.

Relationship between OR and RR

- Recall, from the definition of an **Odds Ratio**

$$\begin{aligned} OR &= \frac{p_1/(1-p_1)}{p_2/(1-p_2)} \\ &= \left(\frac{p_1}{p_2}\right)\left[\frac{1-p_2}{1-p_1}\right] \\ &= RR\left[\frac{1-p_2}{1-p_1}\right] \end{aligned}$$

- When the disease is **rare:**

$$\left[\frac{1-p_2}{1-p_1}\right] \approx \frac{1}{1} = 1; \quad \text{and so} \quad OR \approx RR.$$

- In the Vitamin C example, $\hat{p}_1$ and $\hat{p}_2$ were 0.12 and 0.22, respectively. These are not small enough to be considered rare, and so the estimated OR, $\widehat{OR} = 2.041$ is not that close to the estimated RR, $\widehat{RR} = 1.811$.

- In the example below, aspirin use and heart attacks, $\hat{p}_1$ and $\hat{p}_2$ are both $< 2\%$, so the estimates of the Odds Ratio and Relative Risk are very similar.
  Ex. verify that $\widehat{OR} = 1.832$ and $\widehat{RR} = 1.818$.

Example: Clinical trial for Aspirin Use and Heart Attack in Doctors

		OUTCOME		
		Heart Attack	NO Heart Attack	Total
TREATMENT	Placebo	189	10845	11034
	Aspirin	104	10933	11037
	Total	293	21778	22071

Questions for thought

- Write down the estimates of risk difference, risk ratio, and odds ratio.

- What do we need in order to make inference?

Variance of a treatment difference, in general

- Our treatment differences can be written

$$\theta = g(p_1) - g(p_2).$$

  The estimator is then

$$\hat{\theta} = g(\hat{p}_1) - g(\hat{p}_2)$$

  (Recall that the MLE of $g(\beta)$ is $g(\hat{\beta})$, where $\hat{\beta}$ is MLE of β.)

  Q: What is g in the above for risk difference, risk ratio, and odds ratio?

- Also, since $\hat{p}_1$ and $\hat{p}_2$ are independent (**why**), so are any functions of $\hat{p}_1$ and $\hat{p}_2$. Therefore

$$\text{Var}[g(\hat{p}_1) - g(\hat{p}_2)] = \text{Var}[g(\hat{p}_1)] + \text{Var}[g(\hat{p}_2)]$$

- **Q:** What is $\text{Var}[g(\hat{p}_i)]$?

Example: logit

- We know that
$$\hat{p} = \frac{Y}{n}$$
-
$$\text{Var}(\hat{p}) = \frac{p(1-p)}{n}$$
-
$$g(\hat{p}) = \text{logit}(\hat{p}) = \log\left(\frac{\hat{p}}{1-\hat{p}}\right) = \log(\hat{p}) - \log(1-\hat{p})$$
-
$$\begin{aligned}\frac{\partial[\log(p) - \log(1-p)]}{\partial p} &= \frac{1}{p} - \frac{-1}{1-p} \\ &= \frac{1}{p(1-p)}\end{aligned}$$
- By Delta method, $\text{Var}[g(\hat{p})]$ is approximately
$$\begin{aligned}[g'(\mu)]^2\ \sigma^2 &= \left[\frac{1}{p(1-p)}\right]^2 \frac{p(1-p)}{n} \\ &= \frac{1}{np(1-p)}\end{aligned}$$

The results are summarized in the following table:

Treatment Difference	Estimate	Var(Estimate)
Δ	$\widehat{p}_1 - \widehat{p}_2$	$\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}$
log (RR)	$\log\left(\frac{\widehat{p}_1}{\widehat{p}_2}\right)$	$\frac{1-p_1}{n_1 p_1} + \frac{1-p_2}{n_2 p_2}$
log (OR)	$\log\left(\frac{\widehat{p}_1/(1-\widehat{p}_1)}{\widehat{p}_2/(1-\widehat{p}_2)}\right)$	$\frac{1}{n_1 p_1(1-p_1)} + \frac{1}{n_2 p_2(1-p_2)}$

Ex. Derive these variances if you have not done it before.

We use the log scale, because

- it gets rid of restrictions on the ranges of the parameters, and makes computation a lot easier.
- Also, the **normal approximation** works better.

Confidence Intervals

A 95% **confidence interval** for a parameter β is (L, U), so that

$$P(L < \beta < U) = 0.95,$$

where U and L are calculated from the data. This is the so-called 'interval estimate' ($\hat{\beta}$ is called point estimate).

To find confidence intervals, we often use the fact that for $Y \sim N(\mu, \sigma^2)$,

$$\begin{aligned} & P(-1.96 < \frac{Y - \mu}{\sigma} < 1.96) \\ = \; & P(Y - 1.96\sigma < \mu < Y + 1.96\sigma) = 0.95 \end{aligned}$$

- First of all, we need to estimate the variances – we replace p_1 and p_2 in Var(Estimate) with $\hat{p}_1$ and $\hat{p}_2$.

- Then the 95% confidence intervals for treatment differences can be obtained as

$$(\hat{p}_1 - \hat{p}_2) \ \pm \ 1.96 \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

$$\log(\widehat{RR}) \ \pm \ 1.96 \sqrt{\frac{1-\hat{p}_1}{n_1\hat{p}_1} + \frac{1-\hat{p}_2}{n_2\hat{p}_2}}$$

$$\log(\widehat{OR}) \ \pm \ 1.96 \sqrt{\frac{1}{y_1} + \frac{1}{n_1 - y_1} + \frac{1}{y_2} + \frac{1}{n_2 - y_2}}$$

Ex. verify the last one.

Confidence Intervals for OR and RR

- You might want a confidence interval for RR or OR instead of $\log RR$ or $\log OR$.

- **Q:** what would they be?

Hypothesis Testing

Q: What is the null hypothesis?

Q: What would be a test statistic?

- Under the **null** $\mathrm{H}_0 : p_1 = p_2$, we know that

$$(1) \quad p_1 - p_2 = 0$$

$$(2) \quad \log(RR) = 0$$

$$(3) \quad \log(OR) = 0$$

- Then to test this null hypothesis (against either one or two-sided alternatives), we can use any of the following statistics:

$$(1) \quad Z_1 = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\widehat{\mathrm{Var}}(\hat{p}_1 - \hat{p}_2)}}$$

$$(2) \quad Z_2 = \frac{\log(\widehat{RR})}{\sqrt{\widehat{\mathrm{Var}}[\log(\widehat{RR})]}}$$

$$(3) \quad Z_3 = \frac{\log(\widehat{OR})}{\sqrt{\widehat{\mathrm{Var}}[\log(\widehat{OR})]}}$$

- All three are approximately $N(0,1)$ **under the null**.

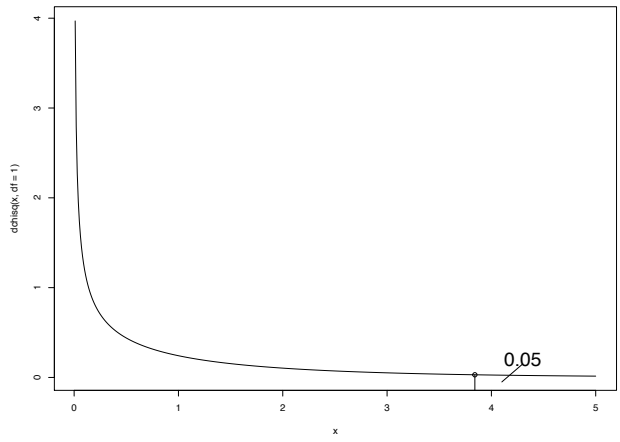
- Suppose that we use Z_i, then for a two-sided test at 0.05 **significance level**, we would reject the null if $|Z_i| > 1.96$

- Note that $|Z_i| > 1.96$ is equivalent to the fact that the corresponding 95% confidence interval does not contain treatment difference 0.

- $Z_1^2 \overset{asymp.}{\sim} \chi_1^2$ under the null gives the *chi-squared test* for a 2×2 contingency table. [R: chisq.test()]

```
> FrenchSkier <-
+ matrix(c(17, 31, 122, 109),
+        nrow = 2,
+        dimnames = list(Treatment = c("V_c", "No V_c"),
+                        Outcome = c("Cold", "No Cold")))

> FrenchSkier
         Outcome
Treatment Cold No Cold
   V_c      17     122
   No V_c   31     109

> chisq.test(FrenchSkier)

Pearson's Chi-squared test with Yates' continuity correction

data:  FrenchSkier
X-squared = 4.1407, df = 1, p-value = 0.04186
```

Q: What is the cutoff for χ^2_1 and significance level 0.05?
Is it one- or two-sided?



Fisher's exact test

		Outcome 1	Outcome 2	
Treatment	1	Y_1	$n_1 - Y_1$	n_1
	2	Y_2	$n_2 - Y_2$	n_2
		$Y_1 + Y_2 = Y$	$n - Y$	$n_1 + n_2 = n$

- When data are sparse, i.e. the *expected* count of any cell is < 5, Fisher's exact test is typically used.
  - Here expected refers to under the null hypothesis that $p_1 = p_2 = p$, $\widehat{E}(Y_i) = n_i\hat{p}$ where $\hat{p} = (Y_1 + Y_2)/(n_1 + n_2)$;
  - Exact inference is based on the fact that, given all 4 marginal totals $(n_1, n_2, Y, n - Y)$ fixed, the first element Y_1 of the contingency table has a *hypergeometric* distribution under the null (Fisher, 1935):

$$P(Y_1 = k) = \frac{\binom{n_1}{k}\binom{n_2}{Y-k}}{\binom{n}{Y}}$$

- R: fisher.test()

```
## Agresti (1990, p. 61f; 2002, p. 91) Fisher's Tea Drinker
## A British woman claimed to be able to distinguish whether milk or
##  tea was added to the cup first.  To test, she was given 8 cups of
##  tea, in four of which milk was added first.  The null hypothesis
##  is that there is no association between the true order of pouring
##  and the woman's guess, the alternative that there is a positive
##  association (that the odds ratio is greater than 1).

TeaTasting <- matrix(c(3, 1, 1, 3), nrow = 2,
       dimnames = list(Guess = c("Milk", "Tea"),
                       Truth = c("Milk", "Tea")))

> TeaTasting
      Truth
Guess  Milk Tea
  Milk    3   1
  Tea     1   3

> fisher.test(TeaTasting)

Fisher's Exact Test for Count Data

data:  TeaTasting
p-value = 0.4857
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
   0.2117329 621.9337505
sample estimates:
odds ratio
  6.408309
```

Q: what is the conclusion here?

$r \times c$ **contingence tables**

Contingence tables can be extended to more than 2 categories for each of the two variables.

- It will no longer be comparing two probabilities.

- There is still chi-squared test for association of the two categorical variables.

- There is also Fisher's exact test when the expected count of any cell is < 5, under the null hypothesis that the two variables are independent.

```
## A r x c table  Agresti (2002, p. 57) Job Satisfaction

Job <- matrix(c(1,2,1,0, 3,3,6,1, 10,10,14,9, 6,7,12,11), 4, 4,
dimnames = list(income = c("< 15k", "15-25k", "25-40k", "> 40k"),
satisfaction = c("VeryD", "LittleD", "ModerateS", "VeryS")))

> Job
        satisfaction
income   VeryD LittleD ModerateS VeryS
  < 15k      1       3        10     6
  15-25k     2       3        10     7
  25-40k     1       6        14    12
  > 40k      0       1         9    11

> fisher.test(Job)

Fisher's Exact Test for Count Data

data:  Job
p-value = 0.7827
alternative hypothesis: two.sided
```

Logistic Regression for a 2×2 table

	Outcome (Y)		
Treatment(X)	1	0	Total
1	Y_1	$n_1 - Y_1$	n_1
0	Y_0	$n_0 - Y_0$	n_0

- Previously, the second row of the table had subscripts of "2", which are now changed to subscripts of "0".

- We considered Y_1 and Y_0 as two separate variables, each of which followed a binomial distribution.

- Now we will define a binary variable X for treatment assignment, with values
  - 0 for placebo or standard treatment
  - 1 for new treatment

- Similarly, we will define a binary variable Y for outcome, with values
  - 0 for failure/no response
  - 1 for success/response

- The number of successes on the new treatment ($X = 1$) is Y_1, with success probability p_1

- The number of successes on the placebo ($X = 0$) is Y_0, with success probability p_0

- We can also write the success probabilities in terms of the conditional probabilities of Y given X:

$$P(Y = 1|X = 1) = p_1$$

$$P(Y = 1|X = 0) = p_0$$

Introduction to Logistic Models

The logistic regression model for the probability of success is

$$\text{pr}[Y = 1|X = x] \quad = \quad p_x \quad = \quad \frac{\exp(\beta_0+\beta_1 x)}{1+\exp(\beta_0+\beta_1 x)}$$

where β_0 and β_1 are parameters, and $x = 0$ or 1. This model can also be written in terms of the logit:

$$\text{logit}(p_x) \quad = \quad \beta_0 + \beta_1 x$$

- If $x = 1$, then

$$\text{pr}[Y = 1|X = 1] = p_1 = \frac{e^{\beta_0+\beta_1}}{1 + e^{\beta_0+\beta_1}}$$

- If $x = 0$, then

$$\text{pr}[Y = 1|X = 0] = p_0 = \frac{e^{\beta_0}}{1 + e^{\beta_0}}$$

- A test of equality of the success probabilities on the two treatments, is therefore equivalent to a test of $\beta_1 = 0$, i.e.,

$$\mathrm{H}_0 : p_1 = p_0 \;\Leftrightarrow\; \mathrm{H}_0 : \beta_1 = 0$$

- We will show that

$$\beta_1 = \log(OR).$$

Properties of the Logistic Regression Model

- The parameters have no restrictions,

$$-\infty < \beta_0 < \infty$$

$$-\infty < \beta_1 < \infty$$

  and for $x = 0, 1$,

$$0 < \left(\frac{e^{\beta_0+\beta_1 x}}{1 + e^{\beta_0+\beta_1 x}} \right) < 1 \qquad \Rightarrow \qquad 0 < p_x < 1$$

- p_x is the probability of success on treatment x, but to compute the odds we also need to know the failure probability:

$$\begin{aligned} \text{pr}[Y = 0|X = x] &= 1 - p_x \\ &= 1 - \frac{e^{\beta_0+\beta_1 x}}{1 + e^{\beta_0+\beta_1 x}} \\ &= \frac{1 + e^{\beta_0+\beta_1 x}}{1 + e^{\beta_0+\beta_1 x}} - \frac{e^{\beta_0+\beta_1 x}}{1 + e^{\beta_0+\beta_1 x}} \\ &= \frac{1}{1 + e^{\beta_0+\beta_1 x}} \end{aligned}$$

The Odds for Treatment x

- The **ODDS** for treatment x ($x = 0$ or 1) is equal to:

$$\begin{aligned} \frac{p_x}{1-p_x} &= \frac{[e^{\beta_0+\beta_1 x}]/[1+e^{\beta_0+\beta_1 x}]}{1/[1+e^{\beta_0+\beta_1 x}]} \\ &= e^{\beta_0+\beta_1 x} \end{aligned}$$

- In many cases, we are interested in the **logit**, or log-odds:

$$\begin{aligned} \text{logit}(p_x) &= \log\left(\frac{p_x}{1-p_x}\right) \\ &= \log\left(e^{\beta_0+\beta_1 x}\right) = \beta_0 + \beta_1 x \end{aligned}$$

- Based on the general formula above, we get:

$$\begin{aligned} \text{logit}(p_1) &= \beta_0 + \beta_1 \cdot 1 = \beta_0 + \beta_1 \\ \text{logit}(p_0) &= \beta_0 + \beta_1 \cdot 0 = \beta_0 \end{aligned}$$

The Odds Ratio from a Logistic Model

- The log-odds ratio is the difference in logits:

$$\begin{aligned} \log(OR) &= \text{logit}(p_1) - \text{logit}(p_0) \\ &= [\beta_0 + \beta_1] - \beta_0 \\ &= \beta_1 \end{aligned}$$

- Equivalently, the **odds ratio** for the new treatment versus placebo ($x = 1$ versus $x = 0$) is therefore:

$$OR = e^{\beta_1}$$

Example: Lung cancer and smoking

- Suppose you are comparing lung cancer and smoking with

$$X = \begin{cases} 1 \text{ if ever smoked (row 1)} \\ 0 \text{ if never smoked (row 2)} \end{cases}$$

$$Y = \begin{cases} 1 \text{ if lung cancer (column 1)} \\ 0 \text{ if no lung cancer (column 2)} \end{cases}$$

- Then we have the following logits for each outcome:

$$\text{logit}(p_1) = \beta_0 + \beta_1$$

$$\text{logit}(p_0) = \beta_0$$

- The log-odds ratio for lung cancer for smokers versus non-smokers is

$$\begin{aligned} \beta_1 &= \text{logit}(p_1) - \text{logit}(p_0) \\ &= \log[OR(\text{smoke} : \text{no smoke})] \end{aligned}$$

Exact test :

Let us consider the problem of testing the value of the parameter p for a binomial random variable with $n = 10$ trials. We wish to test

$$H_0: p = .5$$

versus

$$H_A: p > .5$$

We will use the number of successes, X, as a test statistic; the rejection region will consist of large values of X, those values that are relatively unlikely under H_0 and more likely under H_A. To determine the precise rejection region for a given value of α, we can use this table of cumulative binomial probabilities $[P(X \leq n)]$:

p	0	1	2	3	4	5	6	7	8	9	10
.7	.0000	.0001	.0016	.0106	.0474	.1503	.3504	.6172	.8507	.9718	1.0000
.6	.0001	.0017	.0123	.0548	.1662	.3669	.6177	.8327	.9536	.9940	1.0000
.5	.0010	.0107	.0547	.1719	.3770	.6230	.8281	.9453	.9893	.9990	1.0000

Suppose that the rejection region consists of the points $\{8, 9, 10\}$. The significance level of the test, α, is the probability of rejecting H_0 when it is true; from the last row of the table ($p = .5$), we see that

$$\alpha = P(X > 7) = 1 - P(X \leq 7) = .0547$$

If the rejection region consists of $\{7, 8, 9, 10\}$, the significance level of the test is $\alpha = .172$.

The Neyman–Pearson approach sets a value for α first; suppose that we choose to set $\alpha = .0547$. If the true value of p is .6, the power of the test is the probability that X is greater than or equal to 8; that is, the power is .1673. If the true value is .7, the power is .3828. The power is thus a function of p, and it is not difficult to see that the power tends to 1 as p approaches 1 and that the power tends to α as p approaches .5. □

researcher wishes to test

$$H_0: p = 0.85$$
versus
$$H_1: p \neq 0.85$$

The decision will be based on the magnitude of k, the total number in the sample for whom the durg is effective—that is, on

$$k = k_1 + k_2 + \ldots + k_{19}$$

where

$$k_i = \begin{cases} 0 & \text{if the new drug fails to relieve } i\text{th patient's pain} \\ 1 & \text{if the new drug does relieve } i\text{th patient's pain} \end{cases}$$

What should the decision rule be if the intention is to keep α somewhere near 10%? [Note that Theorem 6.3.1 does not apply here because Inequality 6.3.1 is not satisfied—specifically, $np_o + 3\sqrt{np_o(1 - p_o)} = 19(0.85) + 3\sqrt{19(0.85)(0.15)} = 20.8$ is not less than $n(= 19)$.]

If the null hypothesis is true, the expected number of successes would be $np_o = 19(0.85)$, or 16.2. It follows that values of k to the extreme right or extreme left of 16.2 should constitute the critical region.

```
MTB > pdf;
SUBC > binomial 19 0.85.
```

Probability Density Function

```
Binomial with n = 19 and p = 0.850000

      X        P(X = x)
      8        0.0000
      9        0.0001
     10        0.0007
     11        0.0032      → P(X ≤ 13) = 0.0536
     12        0.0122
     13        0.0374
     14        0.0907
     15        0.1714
     16        0.2428
     17        0.2428
     18        0.1529
     19        0.0456      → P(X = 19) = 0.0456
```

For significance level 0.1, the rejection region is $X<=13$ or $X=19$.

FIGURE 6.3.1

Hypothesis Testing

Basic idea:

1. make (given) an assumption about the underlying distribution of data

   eg. 1) $\mathrm{N}(\mu, \sigma^2)$: $\mu = 1$

   2) two samples from $\mathrm{N}(\mu_1, \sigma^2)$ and $\mathrm{N}(\mu_2, \sigma^2)$: $\mu_1 > \mu_2$

2. look at the data, and decide if the data are consistent with the assumption

This process is formulated as statistical hypothesis testing.

Neyman-Pearson Paradigm

In the framework of statistical hypothesis testing, we make a choice between two mutually exclusive hypotheses:

1. *null* hypothesis H_0, eg. $\mu = 1$, or $\mu_1 = \mu_2$
2. *alternative* hypothesis H_1 (some books use H_A)

- According to H_1, we have
  - one-sided test, eg. H_1: $\mu > 1$, or $\mu_1 < \mu_2$
  - two-sided test, eg. H_1: $\mu \neq 1$, or $\mu_1 \neq \mu_2$

- There is *asymmetry* between H_0 and H_1: we want to answer whether we can reject H_0 or not.

- A decision as to whether or not to reject H_0 in favor of H_1, is made based on the value of a test statistic T.

- The set of values of T for which H_0 is rejected, is called the *rejection region*;

- it is often of the form $T > C$ say, where C is called the *critical value*.

- How to choose C is based on error probabilities.

Two types of error

There are two types of error:

1. type I error: H_0 is true but rejected;
2. type II error: H_1 is true, but H_0 is *not rejected*.

The corresponding error rates are:

- $\alpha = P_{H_0}(\text{“}H_1\text{”}) = P(\text{type I error})$, also called *significance level*, or the *size* of a test;
- $\beta = P_{H_1}(\text{“}H_0\text{”}) = P(\text{type II error})$;
- $1 - \beta = P_{H_1}(\text{“}H_1\text{”})$ is called the *power* of the test.

Truth	Fail to reject H_0	Reject H_0
H_0	no error (1-α)	type I error (α)
H_1	type II error (β)	no error (1-β)

The critical value C is chosen to meet some pre-specified α.

Example

In phase II clinical trials for cancer treatment, we often study the response rate of a drug, i.e. if it shrinks a solid tumor by certain amount. This response rate is often compared to the rate that is achieved by the current standard therapy.

- Data: $X_1, ..., X_n$ from n patients, $X_i = 1$ if patient i achieved response, 0 otherwise.

- Distribution or model: $\sum X_i \sim B(n, p)$

- H_0: $p \leq 0.2$ or $p = 0.2$, where 20% is the response rate achieved by standard therapy

- H_0: $p > 0.2$

- **Q:** what would be a test statistic?

We may use $T = \sum X_i/n = \widehat{p}$

- reject H_0 if $\widehat{p} >$ some value C

- we want $\alpha = P_{p=0.2}(\widehat{p} > C)$

- What is the distribution of $\widehat{p}$ when $p = 0.2$?

$$\widehat{p} \overset{asymp.}{\sim} N(0.2,\, 0.2 \times 0.8/n)$$

- So

$$\begin{aligned} \alpha &= P_{p=0.2}(\widehat{p} > C) \\ &= P\left(\frac{\widehat{p} - 0.2}{\sqrt{0.16/n}} > C_1 = \frac{C - 0.2}{\sqrt{0.16/n}}\right) \\ &\approx P(Z > C_1) \end{aligned}$$

where $Z \sim N(0, 1)$.

- If $\alpha = 0.05$, then $C_1 = 1.65$, because $P(Z > 1.65) = 0.05$.

- Then $C = 0.2 + 0.66/\sqrt{n}$.

- Therefor reject H_0 if $\widehat{p} > 0.2 + 0.66/\sqrt{n}$.

These are mathematical derivations, which provide a *decision rule* that can be applied to any n, and any $\widehat{p}$.

One can think of this as what goes on inside a software.

```
> prop.test(x=40, n=100, p=0.2, alternative="greater")

1-sample proportions test with continuity correction

data:  40 out of 100, null probability 0.2
X-squared = 23.766, df = 1, p-value = 5.44e-07
alternative hypothesis: true p is greater than 0.2
95 percent confidence interval:
 0.3183752 1.0000000
sample estimates:
  p
0.4
```

p-value

Suppose that T is the test statistic, and we reject H_0 if $T > C$.

Now given data $X_1, ..., X_n$, we have calculated the value of T to be T_{obs}. Consider T_{obs} to be a fixed value for now.

p-value $= P_{H_0}(T \geq T_{obs})$.

- p-value is a measure of evidence against H_0; i.e. under H_0, how extreme is the observed data in the direction of H_1.

Eg. (cont'd) Previously we derived the test to reject $H_0 : p = 0.2$ if $\widehat{p} > 0.2 + 0.66/\sqrt{n}$.

Suppose $n = 25$, and $\widehat{p} = 0.3$. **Ex.** do we reject H_0 based on the decision rule above?

Now

$$\begin{aligned} p - \text{value} &= P_{H_0}(\widehat{p} \geq 0.3) \\ &= P\left(Z > \frac{0.3 - 0.2}{0.4/\sqrt{25}} = 1.25\right) \\ &= 0.106 \end{aligned}$$

How exactly do we use p-value?

Note that:

$$\begin{aligned} p - \text{value} < \alpha &\Leftrightarrow P_{H_0}(T \geq T_{obs}) < P_{H_0}(T > C) \\ &\Leftrightarrow T_{obs} > C \\ &\Leftrightarrow \text{reject } H_0 \text{ at } \alpha \text{ level.} \end{aligned}$$

What would be the conclusion of our example?

Note also

- p-value is a random variable, as it is a function of the data;
- it can be shown that, under H_0 the p-value has a Uniform (0, 1) distribution.

Summary of procedure for hypothesis testing

- Based on the research question, set up null and alternative hypotheses, and probabilities of error;
- choose an appropriate statistical model;
- find a test statistic;
- find the null distribution of the test statistic;
- find the rejection region or critical value based on the type I error rate α;
- for data analysis, compute the test statistic or p-value based on data, and decide whether H_0 is rejected;
- for study design, compute the sample size n based on desired power $1 - \beta$.

How to find a test statistic?

- Should have distinguishable values under H_0 versus H_1; eg. tends to be around zero under H_0, and have large values under H_1.

- Need to know its null distribution
  - exact
  - asymptotic

- To develop a test statistic is often a topic of statistical research.

Duality between CI and Hypothesis Test

For a parameter θ in general, $H_0 : \theta = \theta_0$

- for a two-sided alternative $H_1 : \theta \neq \theta_0$, H_0 is rejected if and only if the $100(1-\alpha)\%$ CI does not contain θ_0.

- Similar duality holds for one-sided alternatives. One-sided CI is also referred to as lower/upper confidence bounds:
  - for $H_1 : \theta > \theta_0$, H_0 is rejected if and only if the $100(1-\alpha)\%$ CI of the form (L, ∞) does not contain θ_0.
  - for $H_1 : \theta < \theta_0$, H_0 is rejected if and only if the $100(1-\alpha)\%$ CI of the form $(-\infty, U)$ does not contain θ_0.

Exact Tests

Study questions (refer to the texts for the each of the two examples):

- What are the hypotheses?
- What is the significance level α? one- or two-sided?
- What is the test statistic?
- What is the decision rule corresponding to the α above?
- How is the decision rule derived?
- Can you derive a decision rule for $\alpha = 0.05$?

4.1 Variability in estimates

A natural way to estimate features of the population, such as the population mean weight, is to use the corresponding summary statistic calculated from the sample.[6] The mean weight in the sample of 60 adults in `cdc.samp` is $\bar{x}_{\text{weight}} = 173.3$ lbs; this sample mean is a **point estimate** of the population mean, μ_{weight}. If a different random sample of 60 individuals were taken from `cdc`, the new sample mean would likely be different as a result of **sampling variation**. While estimates generally vary from one sample to another, the population mean is a fixed value.

Guided Practice 4.1 How would one estimate the difference in average weight between men and women? Given that $\bar{x}_{\text{men}} = 185.1$ lbs and $\bar{x}_{\text{women}} = 162.3$ lbs, what is a good point estimate for the population difference?[7]

Point estimates become more accurate with increasing sample size. Figure 4.3 shows the sample mean weight calculated for random samples drawn from `cdc`, where sample size increases by 1 for each draw until sample size equals 500. The red dashed horizontal line in the figure is drawn at the average weight of all adults in `cdc`, 169.7 lbs, which represents the population mean weight.[8]

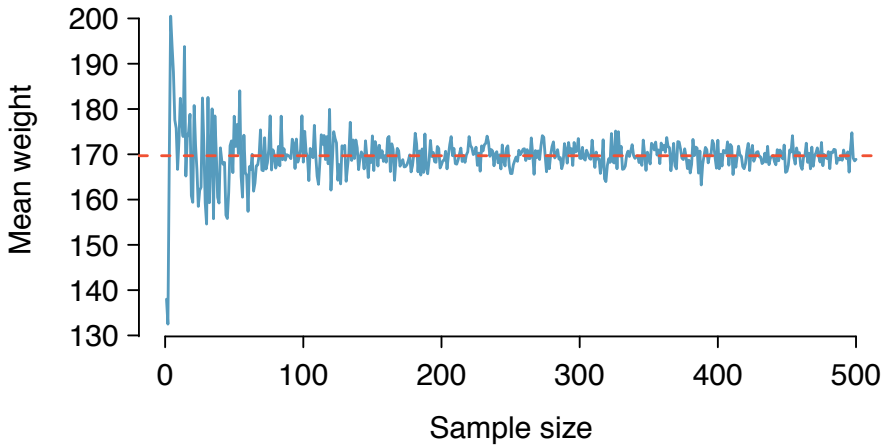


Figure 4.3: The mean `weight` computed for a random sample from `cdc`, increasing sample size one at a time until $n = 500$. The sample mean approaches the population mean (i.e., mean weight in `cdc`) as sample size increases.

Note how a sample size around 50 may produce a sample mean that is as much as 10 lbs higher or lower than the population mean. As sample size increases, the fluctuations around the population mean decrease; in other words, as sample size increases, the sample mean becomes less variable and provides a more reliable estimate of the population mean.

[6] Other population parameters, such as population median or population standard deviation, can also estimated using sample versions.

[7] Given that $\bar{x}_{\text{men}} = 185.1$ lbs and $\bar{x}_{\text{women}} = 162.3$ lbs, the difference of the two sample means, $185.1 - 162.3 = 22.8$lbs, is a point estimate of the difference. The data in the random sample suggests that adult males are, on average, about 23 lbs heavier than adult females.

[8] It is not exactly the mean weight of all US adults, but will be very close since `cdc` is so large.

4.2 Confidence intervals

4.2.1 Interval estimates for a population parameter

While a point estimate consists of a single value, an interval estimate provides a plausible range of values for a parameter. When estimating a population mean μ, a **confidence interval** for μ has the general form

$$(\bar{x} - m,\ \bar{x} + m) = \bar{x} \pm m,$$

where m is the **margin of error**. Intervals that have this form are called **two-sided confidence intervals** because they provide both lower and upper bounds, $\bar{x} - m$ and $\bar{x} + m$, respectively. One-sided sided intervals are discussed in Section 4.2.3.

The standard error of the sample mean is the standard deviation of its distribution; additionally, the distribution of sample means is nearly normal and centered at μ. Under the normal model, the sample mean $\bar{x}$ will be within 1.96 standard errors (i.e., standard deviations) of the population mean μ approximately 95% of the time.[9] Thus, if an interval is constructed that spans 1.96 standard errors from the point estimate in either direction, a data analyst can be 95% **confident** that the interval

$$\bar{x} \pm 1.96 \times \text{SE} \tag{4.2}$$

contains the population mean. The value 95% is an approximation, accurate when the sampling distribution for the sample mean is close to a normal distribution. This assumption holds when the sample size is sufficiently large (guidelines for 'sufficiently large' are given in Section 4.4).

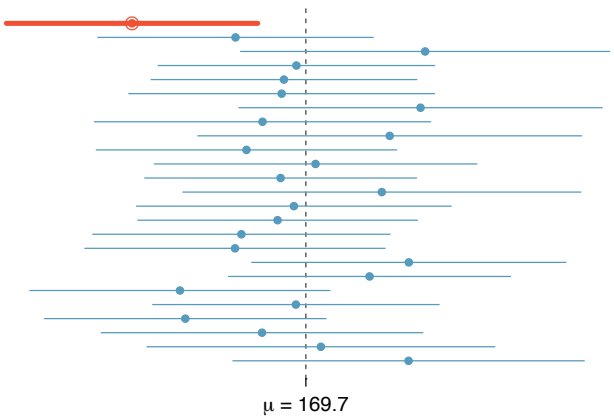


Figure 4.7: Twenty-five samples of size $n = 60$ were taken from cdc. For each sample, a 95% confidence interval was calculated for the population average adult weight. Only 1 of these 25 intervals did not contain the population mean, $\mu = 169.7$ lbs.

The phrase "95% confident" has a subtle interpretation: if many samples were drawn from a population, and a confidence interval is calculated from each one using Equation 4.2, about 95% of those intervals would contain the population mean μ. Figure 4.7

[9] In other words, the Z-score of 1.96 is associated with 2.5% area to the right (and Z = -1.96 has 2.5% area to the left); this can be found on normal probability tables or from using statistical software.

If we denote the interval (L, U), then

$$P(L < \theta < U) = 1-\alpha.$$

(L, U: random; θ: deterministic)

How to construct a c.i.

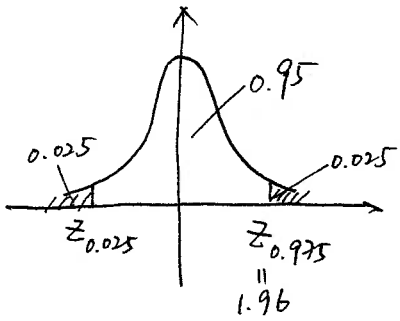
Eg 1. $X_1, \cdots X_n \sim N(\mu, \sigma^2)$ iid, σ^2 known

Fact: $X \sim N(\mu_1, \sigma_1^2)$, $Y \sim N(\mu_2, \sigma_2^2)$ ind

then $X+Y \sim N(\mu_1+\mu_2, \sigma_1^2+\sigma_2^2)$

Exercise: use the above to show that $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$

Then $Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$



$$P(|Z| < 1.96) = 0.95$$

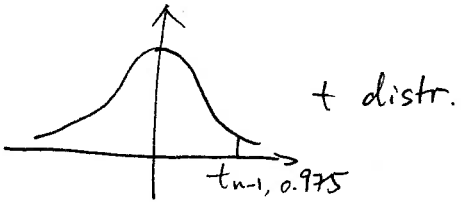
$$= P\left(-1.96 < \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} < 1.96\right)$$

$$= P\left(\bar{X} - 1.96\sigma/\sqrt{n} < \mu < \bar{X} + 1.96\sigma/\sqrt{n}\right)$$

Sometimes write the c.i: $\bar{X} \pm 1.96\sigma/\sqrt{n}$

Eg 2. $X_1, \cdots X_n \sim N(\mu, \sigma^2)$ iid, σ^2 unknown

Fact: $\frac{\bar{X}-\mu}{s/\sqrt{n}} \sim t_{n-1}$



Eg 3. $X_1, \cdots X_n$ iid Poisson (λ) $(EX = \mathrm{Var}X = \lambda)$

then $\dfrac{\bar{X}-\lambda}{\sqrt{\lambda/n}} \to N(0,1)$

To find a 95% c.i. for λ:

① est. λ/n by $\bar{X}/n$, then c.i $\bar{X} \pm 1.96\sqrt{\bar{X}/n}$

eg. study involving asbestos fiber counts

we had $\hat{\lambda} = \bar{X} = 24.9$. $n = 23$

$s_{\hat{\lambda}} = \sqrt{\bar{X}/n} = 1.04$.

$\hat{\lambda} \pm 1.96 s_{\hat{\lambda}}$ is $(22.9, 26.9)$

② $P\left(\left|\dfrac{\bar{X}-\lambda}{\sqrt{\lambda/n}}\right| < 1.96\right) \approx 0.95$

$= P\left(\dfrac{\bar{X} - 2\bar{X}\lambda + \lambda^2}{\lambda/n} < 1.96^2\right)$

$= P\left(\lambda^2 - (2\bar{X} + 1.96^2/n)\lambda + \bar{X} < 0\right)$

solve for L, U.

③ $\lambda > 0$, but previous c.i.'s may contain values < 0

Try to get a c.i. for $\log \lambda$:

$$P(L_1 < \log\lambda < U_1) = 0.95$$

$$= P(e^{L_1} < \lambda < e^{U_1})$$

Delta method

It can be shown that $\log\bar{X} \overset{\text{approx.}}{\sim} N(\log\lambda, \frac{1}{n\lambda})$

then $\dfrac{\log\bar{X} - \log\lambda}{\sqrt{1/n\lambda}} \overset{\text{approx.}}{\sim} N(0,1)$

λ est'd by $\bar{X}$

so 95% c.i for $\log\lambda$: $\log\bar{X} \pm 1.96\sqrt{\frac{1}{n\bar{X}}}$

" " " λ : $\bar{X}e^{\pm \frac{1.96}{\sqrt{n\bar{X}}}}$

④ Try to find a transformation of $\bar{X}$, $g(\bar{X})$, so that Var $g(\bar{X})$ does not depend on λ —— variance stabilizing transformation

Fact: for large n, $\sqrt{\bar{X}} \overset{\text{approx}}{\sim} N(\sqrt{\lambda}, \frac{1}{4n})$.

then another c.i. for λ is

$$\left(\sqrt{\bar{X}} \pm \frac{1.96}{\sqrt{4n}}\right)^2 = \left(\bar{X} + \frac{1.96^2}{4n}\right) \pm 1.96\sqrt{\frac{\bar{X}}{n}}$$

Suppose that we want to study the coverage property of 95% c.i $\bar{X} \pm 1.96\sqrt{\bar{X}/n}$ for λ in the Poisson distribution. (for fixed n, λ)

1) Generate a sample $X_1, \cdots X_n$ from Poisson(λ) distn.

2) Calculate $\bar{X} \pm 1.96\sqrt{\bar{X}/n}$, see if it contains λ

3) Repeat 1) & 2) N times, get the percentage p to the c.i's that contain λ

4) p should be close to 95% (nomial level) in order to conclude that the coverage is good.

Usually we need to do such studies for a variety of values of n, λ & α (significance level)

GENERAL LOGISTIC REGRESSION

So far, we've discussed logistic regression for 2×2 tables as a special case, i.e. the response is binary, and the predictor (regressor) is also binary.

What are possible extensions of the model?

- Continuous covariates as predictors, binary response
  $\Longrightarrow$ multiple logistic regression modeling

- more than 2 response levels ($R \times C$ tables for example)
  - nominal responses (multinomial logistic regression)
  - ordinal responses (ordinal logistic regression)

General Logistic Regression Modeling

We will now extend our logistic regression models to allow multiple covariates, of any type (nominal, ordinal, or continuous)

- In general, we consider a binary response Y_i for the i^{th} individual, and a general vector of covariates:

$$\mathbf{x}_i = [x_{i1}, ..., x_{iK}]'$$

  where x_{ik} is the k^{th} covariate for individual i.

- In the most general case, the x_{ik}'s can represent a combination of both continuous or categorical covariates.

- **ICU Mortality study** (Hosmer & Lemeshow, *Applied Logistic Regression*)

Y	Mortality (died: $Y = 1$, lived: $Y = 0$)
X_1	sex
X_2	age
X_3	level of consciousness (1=normal, 2=stupor, 3=coma)
X_4	race (1=white, 2=black, 3=other)

Another Example – Arthritis Clinical Trial

- This example is from an arthritis clinical trial comparing the drug auranofin to placebo for treatment of rheumatoid arthritis (Bombardier et al., 1986).

- The response of interest is the self-assessment of arthritis, classified as (0) poor or (1) good.

- Individuals were also given a self-assessment at baseline (before treatment), which was also classified as (0) poor or (1) good.

- To make sure that the treatment groups were balanced with respect to baseline status, the randomization occurred after the baseline measurement was taken

- The dataset contains 293 patients who were observed at both baseline and 13 weeks. The data from 25 cases are shown below:

Subset of cases from the arthritis clinical trial

				Self assessment[b]	
CASE	SEX	AGE	TREATMENT[a]	BASELINE	13 WK.
1	M	54	A	0	0
2	M	64	P	0	0
3	M	48	A	1	1
4	F	41	A	1	1
5	M	55	P	1	1
6	M	64	A	1	1
7	M	64	P	1	0
8	F	55	P	1	1
9	M	39	P	1	0
10	F	60	A	0	1
11	M	49	A	0	1
12	M	32	A	0	1
13	F	62	P	0	0
14	M	50	A	0	1
15	M	54	A	0	0
16	M	36	P	1	1
17	M	63	A	1	1
18	F	63	P	0	0
19	M	65	A	1	0
20	M	60	P	1	1
21	F	59	P	1	1
22	M	57	P	1	1
23	M	58	A	0	1
24	F	35	P	1	1
25	F	31	P	0	1

[a] A = Auranofin, P = Placebo

[b] 0=poor, 1=good.

- We are interested in seeing how the binary response

$$Y_i = \begin{cases} 1 \text{ if good at 13 weeks} \\ 0 \text{ if poor at 13 weeks} \end{cases}$$

is affected by the covariates:

1. BASELINE self-assessment:

$$X_i = \begin{cases} 1 \text{ if good at BASELINE} \\ 0 \text{ if poor at BASELINE} \end{cases}$$

2. GENDER

$$\text{SEX} = \begin{cases} 1 \text{ if male} \\ 0 \text{ if female} \end{cases}$$

3. TREATMENT

$$\text{TRT} = \begin{cases} 1 \text{ if auranofin} \\ 0 \text{ if placebo} \end{cases}$$

4. AGE IN YEARS

- The main question is whether the treatment increases the probability of a more favorable response, after controlling for baseline response, age and sex. Secondary questions might be how age and sex affect the probability of response.

Distribution of Response Outcomes

- Since each individual may represent a unique combination of covariates, we no longer count up all those responding within a stratum defined by covariates. Instead, we focus on the distribution of the response for the i^{th} subject:

$$Y_i \sim Bernoulli(p_i)$$

where $p_i = \text{pr}[Y_i = 1|x_{i1}, ..., x_{iK}]$ follows the logistic regression model

$$\text{logit}(p_i) = \log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_{i1} + ... + \beta_K x_{iK}$$

Q: what does i.i.d. refer to?

- We will show in the following that the parameter β_k has the interpretation of a **log-odds ratio** between the response and a one unit increase in the covariate x_{ik}, **conditional** on the other covariates.

- To simplify the interpretation of model parameters, we will temporarily drop the subscript i:

$$\text{logit}(\text{pr}[Y = 1|x_1, ..., x_K]) = \beta_0 + \beta_1 x_1 + ... + \beta_K x_K$$

Interpretation of β_k

- Consider the two logits, where we hold all but x_k constant:

  for $x_k = c$:

  $$\text{logit}(P[Y = 1|x_k = c]) \;=\; \beta_0 + ... + \beta_k c ... + \beta_K x_K$$

  for $x_k = c + 1$:

  $$\text{logit}(P[Y = 1|x_k = c + 1]) \;=\; \beta_0 + ... + \beta_k (c + 1) ... + \beta_K x_K$$

- The log-odds ratio for the two groups is the difference in the logits:

  $$\text{logit}(p|x_k = c + 1) - \text{logit}(p|x_k = c) \;=\; \beta_k$$

- Thus, β_k is the log-odds ratio for a one-unit increase in covariate x_k, given all the other covariates are the same.

- For example, if x_k is a dichotomous covariate which equals 1 for the new treatment and 0 for placebo, then β_k is the log-odds ratio for success for new treatment versus placebo, conditional on the other covariates being the same.

Interaction terms

- Suppose there is an interaction between x_{K-1} and x_K :

$$\text{logit}(p) \;=\; \beta_0 + \beta_1 x_1 + ... + \beta_{K-1} x_{K-1} + \beta_K x_K + \gamma x_{K-1} x_K$$

- Now, if we compare the same two logits as before:

$$\begin{aligned}
&\text{logit}(p|x_K = c+1) - \text{logit}(p|x_K = c) \\
&\quad = \beta_0 + \beta_1 x_1 + ... + \beta_{K-1} x_{K-1} + \beta_K (c+1) + \gamma x_{K-1}(c+1) \\
&\qquad - \beta_0 + \beta_1 x_1 + ... + \beta_{K-1} x_{K-1} + \beta_K c + \gamma x_{K-1} c \\
&\quad = \beta_K + \gamma x_{K-1}
\end{aligned}$$

- Thus, conditional on the first $(K-1)$ covariates, the log-odds ratio for a one unit increase in the K^{th} covariate is

$$\beta_K + \gamma x_{K-1}$$

  and **depends on the level of** x_{K-1}

- We could include both two-way and three-way interactions, but interpretation of interactions terms becomes complicated even with just two-way interactions.

Main effects model results

Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Wald Chi-Square	Pr > Chi-Square	Standardized Estimate	Odds Ratio
INTERCPT	1	0.3327	0.8409	0.1566	0.6923	.	1.395
SEX	1	0.2168	0.3389	0.4095	0.5222	0.053354	1.242
AGE	1	-0.00530	0.0144	0.1361	0.7122	-0.032426	0.995
TRT	1	0.7005	0.3136	4.9897	0.0255	0.193432	2.015
X	1	1.4231	0.3102	21.0539	0.0001	0.365832	4.150

Q: what is the meaning of intercept?

Maximum Likelihood Estimation (MLE) for Logistic Regression

- Consider the general logistic regression model

$$\text{logit}(p_i) \;=\; \beta_0 + \beta_1 x_{i1} + ... + \beta_K x_{iK}$$

$$\text{with} \quad Y_i \;\sim\; Bernoulli(p_i) \qquad i = 1, ..., n$$

- The **likelihood**, i.e. probability of observed data, is

$$L(\beta_0, \beta_1, ..., \beta_K) = \prod_{i=1}^{n} p_i^{y_i}(1 - p_i)^{(1-y_i)}$$

where

$$p_i = \frac{\exp(\beta_0 + \beta_1 x_{i1} + ... + \beta_K x_{iK})}{1 + \exp(\beta_0 + \beta_1 x_{i1} + ... + \beta_K x_{iK})}.$$

- The **log-likelihood** is

$$\log[L(\beta_0, \beta_1, ..., \beta_K)]$$

$$= \beta_0 \left(\Sigma_{i=1}^{n}\, y_i\right) + \beta_1 \left(\Sigma_{i=1}^{n}\, x_{i1} y_i\right) + ... + \beta_K \left(\Sigma_{i=1}^{n}\, x_{iK} y_i\right)$$

$$- \Sigma_{i=1}^{n} \log(1 + e^{\beta_0 + \beta_1 x_{i1} + ... + \beta_K x_{iK}})$$

- The **MLE** is the value of $\beta_0, \beta_1, ..., \beta_K$ that maximizes the (log-)likelihood function.

- It is possible that the data are so sparse (most people respond 0 or most people respond 1) that there is no solution to the maximization problem. (This is different from linear regression.)

- But if there is a solution, it can be shown to be unique.

- In practice, if there is no solution, your logistic regression software will say something like 'Convergence not reached after xx iterations'.

- The smaller of the number of 0's or number of 1's, call it the number of 'events", is the **'effective' sample size** for logistic regression. (*Rule of thumb* says that you need at least 10 events for each parameter to be estimated.)

- There are other methods that may be more appropriate with sparse data (conditional logistic regression, exact methods).

Confidence Intervals

95% (asymptotic) confidence interval for β_k can be obtained via

$$\widehat{\beta}_k \pm 1.96 \sqrt{\widehat{\text{Var}}(\widehat{\beta}_k)}$$

- In software outputs, you can look under the column labeled "standard error" to get the square root of the variance for a particular parameter estimate.

- In fact the estimated variance-covariance matrix for the entire vector of estimates $\widehat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_K)$ is also given by the statistical software packages.

Confidence interval for a linear combination

- Suppose we want a 95% confidence intervals for a linear combination (sometimes called 'contrast') of $\boldsymbol{\beta}$ of the form

$$\mathbf{c}\boldsymbol{\beta} = c_0\beta_0 + c_1\beta_1 + ... + c_K\beta_K$$

for some set of constants $\mathbf{c} = [c_0, c_1, ..., c_K]$

- For example, consider a model with interaction

$$\text{logit}(p_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_{12} x_{i1} x_{i2}$$

- The log-odds ratio for a one unit increase in x_{i2}, given a value of $x_{i1} = x_1$ is

$$\beta_2 + \beta_{12} x_1$$

- For this model, the parameter vector and contrast of interest are:

$$\boldsymbol{\beta}' = [\beta_0, \beta_1, \beta_2, \beta_{12}]$$

$$\mathbf{c} = [0 \;\; 0 \;\; 1 \;\; x_1]$$

- For a 95% confidence intervals for $\mathbf{c}\beta$, we would use

$$\mathbf{c}\widehat{\boldsymbol{\beta}} \pm 1.96 \sqrt{\mathbf{c}\widehat{\text{Var}}[\widehat{\boldsymbol{\beta}}]\mathbf{c}'}$$

$$\text{since} \quad \text{Var}[\mathbf{c}\widehat{\boldsymbol{\beta}}] = \mathbf{c}\text{Var}[\widehat{\boldsymbol{\beta}}]\mathbf{c}'$$

Test Statistics for parameters β_k

- For the logistic model, a test of $H_0 : \beta_k = 0$ represents a test of whether the k^{th} covariate (x_{ik}) affects the probability of success, with the null hypothesis that the probability of success is independent of x_{ik}.

- **Wald Statistic:**

$$Z = \frac{\widehat{\beta}_k}{\sqrt{\widehat{\mathrm{Var}}(\widehat{\boldsymbol{\beta}})_{k+1,k+1}}}$$

This test statistic is asymptotically $N(0,1)$ under the null in large samples.

- **Likelihood Ratio Statistic:**

$$\begin{aligned} T &= 2\{\log L(\widehat{\beta}_0, \widehat{\beta}_1, ..., \widehat{\beta}_K) \\ &\qquad - \log L(\tilde{\beta}_0, \tilde{\beta}_1, ..., \tilde{\beta}_k = 0, ..., \tilde{\beta}_K | H_0)\} \\ &= 2 \sum_{j=1}^{n} \left[y_i \log \left(\frac{\widehat{p}_i}{\tilde{p}_i} \right) + (1 - y_i) \log \left(\frac{1 - \widehat{p}_i}{1 - \tilde{p}_i} \right) \right] \end{aligned}$$

where $\widehat{p}_j$ is the MLE, and $\tilde{p}_j$ is the estimate under the null (remember p is a function of the β's). This test statistic follows a χ^2_1 distribution under the null in large samples.

Score test statistic (sometimes called "**Rao's**")

- The score test statistic is:

$$X^2 = \frac{[\Sigma_{i=1}^n x_{ik}(y_i - \tilde{p}_i)]^2}{\widehat{\text{Var}}[\Sigma_{i=1}^n x_{ik}(y_i - \tilde{p}_i)]}$$

- It is computed under the null, and hence $\tilde{p}_i$ in the expression.

Likelihood Ratio Test for Nested Models

- Sometimes you have nested models resulting from, for example, putting additional interaction terms and/or square terms in the model and testing their significance.

- For example, suppose you have Model 1,

  Model 1:

$$p_i = \frac{e^{\beta_0 + \boldsymbol{\beta}_1' \mathbf{x}_i}}{1 + e^{\beta_0 + \boldsymbol{\beta}_1' \mathbf{x}_i}}$$

- This model is nested in Model 2:

  Model 2:

$$p_i = \frac{e^{\beta_0 + \boldsymbol{\beta}_1' \mathbf{x}_i + \boldsymbol{\beta}_2' \mathbf{z}_i}}{1 + e^{\beta_0 + \boldsymbol{\beta}_1' \mathbf{x}_i + \boldsymbol{\beta}_2' \mathbf{z}_i}}$$

- We want to test

$$H_0 : \boldsymbol{\beta}_2 = 0$$

- The model with more parameters will always have a larger value for the maximized likelihood, since it is maximized over a larger parameter space.

- The difference between the maximized log likelihoods (i.e. likelihood ratio statistic) can be used to test for significance of the extra parameters in model 2 versus model 1:

$$\Delta = 2\{\log L(\hat{\boldsymbol{\beta}}|\mathrm{M}_2) - \log L(\tilde{\boldsymbol{\beta}}|\mathrm{M}_1)\}$$

- If the smaller model fits, i.e. under the null, Δ follows a χ^2_m distribution in large samples, where m parameters are set to 0 in the smaller model.

Fitted example: ICU data

This data set is available in R package 'aplore3'.

```
> summary(icu$sta)
Lived  Died
  160    40

icu.fit <- glm(sta ~ gender + age + race + loc, family = binomial(),
data = icu)

summary(icu.fit)

Call:
glm(formula = sta ~ gender + age + race + loc, family = binomial(),
    data = icu)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.03430  -0.63669  -0.51957  -0.00017   2.35145

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)    -3.19021    0.82094  -3.886 0.000102 ***
genderFemale   -0.16447    0.42932  -0.383 0.701652
age             0.02582    0.01232   2.096 0.036048 *
raceBlack     -16.26983 1464.46752  -0.011 0.991136
raceOther      -0.13027    1.10471  -0.118 0.906131
locStupor      34.91061 2822.61314   0.012 0.990132
locComa         2.99064    0.82556   3.623 0.000292 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

```
(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 200.16  on 199  degrees of freedom
Residual deviance: 153.91  on 193  degrees of freedom
AIC: 167.91

Number of Fisher Scoring iterations: 17
```

Note that residual deviance here is -2 times the log likelihood evaluated at the MLE.

Q: is this a good model to fit?

```
> summary(icu$race)
White Black Other
  175    15    10
> summary(icu$loc)
Nothing  Stupor    Coma
    185       5      10
```

- Testing nested models:

```
> anova(icu.fit, test = "LRT")
Analysis of Deviance Table

Model: binomial, link: logit

Response: sta

Terms added sequentially (first to last)


       Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                     199     200.16
gender  1    0.084       198     200.08  0.771314
age     1    7.771       197     192.31  0.005309 **
race    2    1.256       195     191.05  0.533636
loc     2   37.140       193     153.91 8.614e-09 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

Q: what are the values of the likelihood ratio test statistic?

More on Logistic Regression

Under the model for subject i

$$p_i = \frac{e^{\beta_0 + \boldsymbol{\beta}_1' \mathbf{x}_i}}{1 + e^{\beta_0 + \boldsymbol{\beta}_1' \mathbf{x}_i}}.$$

- The estimated 'risk score' is $\widehat{\beta}_0 + \widehat{\boldsymbol{\beta}}_1' \mathbf{x}_i$, which is the log-odds, also referred to as the linear predictors.

- The estimated or predicted probabilities of response is

$$\hat{p}_i = \frac{e^{\widehat{\beta}_0 + \widehat{\boldsymbol{\beta}}_1' \mathbf{x}_i}}{1 + e^{\widehat{\beta}_0 + \widehat{\boldsymbol{\beta}}_1' \mathbf{x}_i}}.$$

```
> summary(predict(icu.fit))
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-19.005  -2.160  -1.625  -1.971  -1.357  33.657

> summary(predict(icu.fit, type = "response"))
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.0000  0.1034  0.1646  0.2000  0.2048  1.0000
```

Earlier there was warning:

```
> icu.fit <- glm(sta ~ gender + age + race + loc,
family = binomial(), data = icu)

Warning message:
glm.fit: fitted probabilities numerically 0 or 1 occurred
```

How to Build a Model

- Knowledge of the subject matter area, obtaining data.

- We then typically start by looking at the data, literally, as we discussed before.

- Descriptive summary statistics such as those in 'Table 1' are often calculated.

- When we have an outcome variable of interest, in this case binary, we first examine the relationship between it and *each* of the other variables that are potentially predictors.
  - This is to **screen** out the variables that might be considered 'noise';
  - Noise variables can often impact the final performance of the model in a negative way.

  Eg. in the SEER-Medicare data example we showed at the beginning of this course (Hou *et al.*, 2018), from near 9,000 insurance claims codes, we screened down to 2,188 codes for non-cancer mortality, and 1,079 codes for cancer mortality. Screening is also commonly done for -omics studies.

  Q: how do you examine this relationship?

- We then consider multivariate (some call 'multivariable') logistic regression models in this case.

  Q: what is the purpose of building a model?

  [Hint] consider the homework assignment about prediction and causal inference.

Tools for Model Selection

There are various approaches to model selection. In practice, model selection can be done through a combination of them.

- Stepwise procedures
  - forward selection
  - backward selection
  - stepwise selection

- Measures of explained variation, eg. R^2

- Information criteria, eg. AIC, BIC, etc.

- Other dimension reduction methods for high-dimensional data.

Stepwise Procedures

Stepwise procedures have been criticized for being *ad hoc* etc, but continue to be very widely used in practice.

Caution: 'stepAIC' in R gives strange results, and is not recommended. (They are available as automated procedures in Stata and SAS.)
R package 'SignifReg' has p-value (see below) as criterion but only for linear regression.

We briefly describe the stepwise (back-n-forth) procedure there:

(1) Fit a univariate model for each covariate, and identify the predictors significant at some level p_1, say 0.20. (This is the screening step.)

(2) Fit a multivariate model with all significant univariate predictors, and use *backward* selection to eliminate non-significant variables at some level p_2, say 0.10.

(3) Starting with final step (2) model, consider each of the non-significant variables from step (1) using *forward* selection, with significance level p_3, say 0.10.

(4) Do final pruning of main-effects model (omit variables that are non-significant, add any that are significant), using *stepwise* regression with significance level p_4. At this stage, you may also consider adding interactions between any of the main effects currently in the model.

An illustration example:

Survival of Atlantic Halibut (Smith *et al.*)

Obs #	*Surv*ival *Time* (min)	Survival Indicator	*Tow Dur*ation (min.)	Diff in *Depth*	*Length* of Fish (cm)	*Handling* Time (min.)	Total *log(catch)* ln(weight)
100	353.0	1	30	15	39	5	5.685
109	111.0	1	100	5	44	29	8.690
113	64.0	0	100	10	53	4	5.323
116	500.0	1	100	10	44	4	5.323
⋮							

The following is results of Forward Selection in Stata, using p-value < 0.05 as entry criterion.

```
                    begin with empty model

p = 0.0000 <  0.0500  adding   handling
p = 0.0000 <  0.0500  adding   logcatch
p = 0.0010 <  0.0500  adding   towdur
p = 0.0003 <  0.0500  adding   length
```

```
------------------------------------------------------------------------------
survtime |
  censor |      Coef.   Std. Err.       z     P>|z|     [95% Conf. Interval]
---------+--------------------------------------------------------------------
handling |   .0548994   .0098804      5.556   0.000     .0355341    .0742647
logcatch |  -.1846548    .051015     -3.620   0.000     .2846423   -.0846674
  towdur |   .5417745   .1414018      3.831   0.000     .2646321     .818917
  length |  -.0366503   .0100321     -3.653   0.000    -.0563129   -.0169877
------------------------------------------------------------------------------
```

The following is results of Backward Selection in Stata, using p-value ≥ 0.05 as removal criterion.

```
                    begin with full model

p = 0.1991 >= 0.0500  removing depth


------------------------------------------------------------------------------
survtime |
  censor |      Coef.   Std. Err.       z     P>|z|   [95% Conf. Interval]
---------+--------------------------------------------------------------------
  towdur |   .5417745   .1414018      3.831   0.000   .2646321     .818917
logcatch |  -.1846548    .051015     -3.620   0.000  -.2846423   -.0846674
  length |  -.0366503   .0100321     -3.653   0.000  -.0563129   -.0169877
handling |   .0548994   .0098804      5.556   0.000   .0355341    .0742647
------------------------------------------------------------------------------
```

The following is results of Stepwise Selection in Stata, using p-value < 0.05 as entry criterion, and p-value ≥ 0.10 as removal criterion.

```
                    begin with full model

p = 0.1991 >= 0.1000  removing depth

------------------------------------------------------------------------------
survtime |
  censor |      Coef.   Std. Err.       z     P>|z|  [95% Conf. Interval]
---------+--------------------------------------------------------------------
  towdur |   .5417745   .1414018      3.831   0.000  .2646321     .818917
handling |   .0548994   .0098804      5.556   0.000  .0355341    .0742647
  length |  -.0366503   .0100321     -3.653   0.000 -.0563129   -.0169877
logcatch |  -.1846548    .051015     -3.620   0.000 -.2846423   -.0846674
------------------------------------------------------------------------------
```

Notes:

- When the halibut data was analyzed with the forward, backward and stepwise options, the same final model was reached. However, this will not always be the case.

- Sometimes we want to force certain variables in the models during the whole selection process, even if they may not be significant.

- Depending on the software, different tests (Wald, score, or likelihood ratio) may be used to decide what variables to add and what variables to remove.

Interactions

It is always a good idea to check for interactions:

In this example, there are several important interactions. Here backward selection was used, while forcing all main effects to be included, and considering all pairwise interactions. Here are the results:

Variable	DF	Parameter Estimate	Standard Error	Wald Chi-Square	Pr > Chi-Square	Exp of Estimate
TOWDUR	1	-0.075452	0.01740	18.79679	0.0001	0.927
DEPTH	1	0.123293	0.06400	3.71107	0.0541	1.131
LENGTH	1	-0.077300	0.02551	9.18225	0.0024	0.926
HANDLING	1	0.004798	0.03221	0.02219	0.8816	1.005
LOGCATCH	1	-0.225158	0.07156	9.89924	0.0017	0.798
TOWDEPTH	1	0.002931	0.0004996	34.40781	0.0001	1.003
TOWLNGTH	1	0.001180	0.0003541	11.10036	0.0009	1.001
TOWHAND	1	0.001107	0.0003558	9.67706	0.0019	1.001
DEPLNGTH	1	-0.006034	0.00136	19.77360	0.0001	0.994
DEPHAND	1	-0.004104	0.00118	12.00517	0.0005	0.996

Interpretation:

Handling alone doesn't seem to affect survival, unless it is combined with a longer towing duration or shallower trawling depths.

Hierarchical principle: if an interaction is included in a model, then the main effects are included.

An alternative modeling strategy when we have fewer covariates

With a dataset with only 5 main effects, you might be able to consider interactions from the start. How many would there be?

- Fit model with all main effects and pairwise interactions
- Then use backward selection to eliminate non-significant pairwise interactions (remember to force the main effects into the model at this stage, according to the 'hierarchical principle')
- Once non-significant pairwise interactions have been eliminated, you could consider backwards selection to eliminate any non-significant main effects that are not involved in remaining interaction terms

R^2-type Measures

R^2-type measures have always been considered useful in practice, to quantify how much variation in the outcome is explained by the regressors or predictors.

More recently:

- It has also been used to quantify genetic heritability, including the *polygenic risk scores*.
- For predictability, *out of sample* R^2 has been used in machine learning approaches.

Eg. In a prognostic study in gastric cancer, we wanted to investigate the prognostic effects of blood-based acute phase reactant proteins (i.e. *biomarkers*) and stage on survival. Note that stage is only available after surgery. The types of questions we were interested in:

1. How much of the variability in survival is explained, by the biomarkers and/or stage?

2. How strong are the effects of certain prognostic variables once others have been accounted for?

3. How much predictability is lost, if at all, when replacing a continuously measured covariate by a binary coding?

4. In some other disease areas, eg. CD4 counts in AIDS patients, how much is the effect on survival "captured" by such a *surrogate*?

Note that the R^2 measure concerns explained variation, or predictive capability, but **not** the goodness-of-fit of a model (which is a common misunderstanding).

Counter example: in linear regression, if the regression line is flat i.e. slope is close to zero, then R^2 is close to zero. But the fit can be good. The regressors just have little predictive power.

R^2 measure for logistic regression

For linear regression the R^2 measure, also called the coefficient of determination, is well-known. It is the proportion of the variance in the dependent variable that is explained by the independent variable(s).

For logistic regression (or binary outcome in general), a generalized R^2 (Cox and Snell) can be easily calculated using the likelihood ratio statistic:

- The measure can be defined as

$$R^2 = 1 - e^{-\Gamma},$$

  where

$$\Gamma \;=\; 2\{\log L(\hat{\boldsymbol{\beta}}) - \log L(\mathbf{0})\}/n,$$

  which is the likelihood ratio test statistic divided by n the sample size.

- It is
  - between 0 and 1 (why);
  - if $\hat{\boldsymbol{\beta}} = \mathbf{0}$, then $R^2 = 0$, therefore no regression effect translates to R^2 that is very close to zero;
  - increasing R^2 values generally indicate increasing predictability of the model (i.e. the regressors);

- for nested models, the R^2 value is non-decreasing with the larger model(s) (why);
- the use of R^2 can often be framed as: does the inclusion of additional predictors lead to substantial increase in R^2?

- R^2 can be seen as the **proportion of the explained randomness** (Kent, 1983), which is related to the Kullback-Leibler information:

$$R^2 = 1 - \frac{D(\hat{\boldsymbol{\beta}})}{D(\mathbf{0})},$$

where $D(\mathbf{0}) = \exp\{-2\log L(\mathbf{0})/n\}$ is the *randomness* in Y, and $D(\hat{\boldsymbol{\beta}}) = \exp\{-2\log L(\hat{\boldsymbol{\beta}})/n\}$ is the *residual randomness* of Y explained by $\mathbf{X}$.

- This mimics explained variation (under linear regression), which would be

$$1 - \frac{E\{\mathrm{Var}(Y|\mathbf{X})\}}{\mathrm{Var}(Y)} = \frac{\mathrm{Var}\{E(Y|\mathbf{X})\}}{\mathrm{Var}(Y)}.$$

- Γ estimates twice the Kullback-Leibler information gain, between the fitted model and the null model.

Information Criteria

Information criteria have been used for model selection.

Akaike Information (AI):

- *Risk functions* are often used to evaluate a model, or a statistical procedure in general. Typically the smaller the risk the better.

- Risks are often defined as the expected value of a *loss function*.
  - Eg. squared error loss gives rise to mean squared error (MSE) as a risk function:
  - for estimating a population parameter θ and any estimator $\hat{\theta}$, the squared error is $(\hat{\theta} - \theta)^2$, so

$$\text{MSE}(\hat{\theta}) = E(\hat{\theta} - \theta)^2 = \text{Var}(\hat{\theta}) + \text{bias}(\hat{\theta})^2,$$

    where $\text{bias}(\hat{\theta}) = E(\hat{\theta}) - \theta$. [Ex.]

- For AI we consider the *deviance loss* function $l(y, \theta) = -2 \log g_\theta(y)$, where g_θ is the density (or probability function) of a family of distributions indexed by parameter θ, that is used to model the observed data y.
  - Eg. the family might be Bernoulli, with a logistic regression model $\theta = (\beta_0, \beta_1, ..., \beta_K)'$.

- The corresponding risk function is $E\{-2\log g_\theta(y)\}$, closely related to the *Kullback-Leibler (KL) information* $E\{\log g_\theta(y)\}$.

- We want to choose the model that minimizes the above risk.

- Note that the KL information gain gives a kind of 'distance' from the true distribution f that generates the data y, to g_θ:
$$KL(f, g_\theta) = E_f\{\log f(y) - \log g_\theta(y)\}.$$

- For a given family g_θ, minimum KL is attained at θ_0 such that $KL(f, g_{\theta_0}) = \min_\theta KL(f, g_\theta)$ or, equivalently,
$$E\{\log g_{\theta_0}(y)\} = \max_\theta E\{\log g_\theta(y)\}.$$

- When the model is *correct*, we have $f = g_{\theta_0}$.

- In practice θ_0 is often estimated by the MLE $\hat{\theta}(y)$.

- Then the risk $-2E\{\log g_{\theta_0}(y)\}$ is 'estimated' by
$$-2E_{y^*}\{\log g(y^*|\hat{\theta}(y))\}.$$
Note that we use y^* to denote the r.v. that the expectation is w.r.t., in order to distinguish from y the observed data that's used to estimate θ_0.

- The **expected risk** in this case is the Akaike Information:

$$AI = -2E_y E_{y^*}\{\log g(y^*|\hat{\theta}(y))\}. \quad (1)$$

  It is also referred to as the *predictive* log-likelihood, or the expected KL. Note that y^* is an independent replicate of y, i.e. from the same distribution as y.

- The model should be chosen to minimize the AI, which itself needs to be estimated.

- **Q:** how would you estimate AI?

- It has been known that the 'apparent' estimate $-2 \log g(y|\hat{\theta}(y))$ under-estimates AI. (why)

- Instead Akaike (1973) showed that

$$AIC = -2 \log g(y|\hat{\theta}(y)) + 2p \qquad (2)$$

  is an approximately unbiased estimator of AI, where p is the dimension of θ.

- Therefore the model is chosen to minimize the AIC.

See ICU example for the computed AIC value.

Bayesian information criterion (BIC)

If p is the number of parameter in a model, the Bayesian information criterion is

$$BIC = -2 \log g(y|\hat{\theta}(y)) + p \cdot \log(n), \tag{3}$$

where n is the sample size.

Penalized log-likelihood

Almost all the methods for model selection we discuss here can be written as choosing β to maximize a penalized log-likelihood:

$$\log g(y|\beta) - P_\lambda(\beta),$$

where $\lambda \geq 0$ is the penalty parameter, and often we can use the penalty $P_\lambda(\beta) = \lambda \Sigma_{j=1}^{p} |\beta_j|^m$.

1. $m = 0$, L_0 penalty: best subset (AIC), stepwise (might require orthonormal design under linear regression), adjusted R^2, generalized cross-validation (GCV).

2. $m = 1$, L_1 penalty: least absolute shrinkage and selection operator (LASSO).

3. $m = 2$, L_2 penalty: ridge regression.

4. Other penalties: elastic net (combined L_1 and L_2 penalties), smoothly clipped absolute deviation (SCAD) etc.

See Harezlak et al. Chapter in "*High-Dimensional Data Analysis in Cancer Research*", Springer 2009.

Other Regression Models for Binary Outcome

- Although logistic regression is by far the most popular way to model Bernoulli data, we can also use other link functions.

- Since a probability must be between 0 and 1, we would like to model

$$p_i = \text{pr}[Y_i = 1 | x_{i1}, ..., x_{iK}]$$

as a function of covariates and parameters that will always be between 0 and 1.

- In **logistic regression**:

$$p_i = F(\beta_0 + \beta_1 x_{i1} + ... + \beta_K x_{iK}) = \frac{e^{\beta_0 + \beta_1 x_{i1} + ... + \beta_K x_{iK}}}{1 + e^{\beta_0 + \beta_1 x_{i1} + ... + \beta_K x_{iK}}}$$

where

$$F(u) = \text{pr}[U \leq u] = \frac{e^u}{1 + e^u}$$

is the cumulative distribution function (CDF) of the logistic distribution.

- In general, we can use any CDF to model p_i, since

$$F(u) = \text{pr}[U \leq u] \in [0, 1]$$

- So we can model

$$p_i = F(\beta_0 + \beta_1 x_{i1} + ... + \beta_K x_{iK})$$

where $F(u)$ is any CDF.

- You can think of $F(u)$ as a function that maps a number

$$-\infty < u < \infty \quad \Longrightarrow \quad 0 < F(u) < 1$$

- The nice thing about this structure is that it allows us to model $F^{-1}(p_i)$ as a **linear** function of the covariates:

$$F^{-1}(p_i) = \mathbf{x}'\boldsymbol{\beta} = \beta_0 + \beta_1 x_{i1} + ... + \beta_K x_{iK}$$

- For example, for the logistic link: $F^{-1}(p_i) = \text{logit}(p_i)$

Some examples of link functions:

(1) The logistic link

(2) The Complementary log-log link

(3) The Probit link

Complementary log-log Link

- The CDF from the extreme value distribution is

$$F(u) = \exp[-\exp(-u)]$$

- If we substitute this CDF in $F(\mathbf{x}'\boldsymbol{\beta})$, we get:

$$p_i = \exp[-\exp(\beta_0 + \beta_1 x_{i1} + ... + \beta_K x_{iK})]$$

- This is equivalent to modeling the transformation $\log(-\log(p_i))$ as:

$$\log[-\log(p_i)] = \beta_0 + \beta_1 x_{i1} + ... + \beta_K x_{iK}$$

  which is why it is called the complementary log-log link.

- It is often used for discrete survival data (when thought of as discretizing an underlying latent random variable that follows a proportional hazards model).

The Probit Link

- The **Probit** link corresponds to the standard normal $N(0,1)$ CDF

$$F(u) = \int_{-\infty}^{u} e^{-\frac{u^2}{2}} du = \Phi(u)$$

- There is no 'closed form expression' for $\Phi(u)$ as there is in the logistic, so we usually just denote it by $\Phi(u)$. For a given value of u, you use the computer to find $\Phi(u)$.

- The probit model can therefore be expressed as:

$$p_i = \Phi(\beta_0 + \beta_1 x_{i1} + ... + \beta_K x_{iK})$$

$$\text{or as } \Phi^{-1}(p_i) = \beta_0 + \beta_1 x_{i1} + ... + \beta_K x_{iK}$$

- The fit of models using a probit link is typically very similar to the logistic link.

- This model has a biological interpretation related to tolerance distributions, and is therefore used for modeling data from dose-response studies. Some dose-response studies involve the identification of a "threshold" dose, above which you get the response ($Y = 1$) and below which you do not ($Y = 0$).

Maximum Likelihood Methods for Other Links

- Just as we showed for logistic regression, maximum likelihood methods can be used to estimate the parameters of these models, i.e., maximize

$$L(\beta) = \prod_{i=1}^{n} p_i^{y_i}(1-p_i)^{1-y_i}$$

- The MLE can be obtained by setting the first derivative vector of the log likelihood to $\mathbf{0}$ and solving for $\widehat{\boldsymbol{\beta}}$, and the negative inverse of the estimated second derivative matrix can be used to estimate the variance.

- The solution is usually obtained by the Newton-Raphson algorithm, just as in logistic regression. You can use SAS Proc Logistic or Proc Probit.