1 Introduction

Outline

1. Lecture in person
2. Office Hours in person or can be arranged to be online using Google Meet.
3. Homework/Labwork

   Lecture and lab assignments will be combined into one assignment due every other week, use the assignments to convince yourself that you have learned the material

   Groups of up to three students will complete a big major data analytic project. Students will be expected to meet online or in person to complete the project. The final report will be a group submission.
4. Web site: link in Meskanas, has online notes and extra resources, not all will be copied into Meskanas
5. Email: use email to contact me, we can arrange e-chats if necessary
6. Textbook (recommended): Agresti "Introduction to Categorical Data Analysis"

   I really like this book, but it is quite expensive. My online notes align with the presentation in the book and use the same notation. The book provides additional explanations of concepts and ideas, and always presents extra examples.
7. Check out the Meeting Room in the Math & Stats Department. Use it as a study space or find class mates.

Pre-reqs

- Assumption, you can interpret statistics:
  - What does it mean to be 95% confident?
  - What is inferential statistics?

  From STAT 151, STAT 252, STAT 266:
  - What is measured by the P-value?
  - Rational in testing
  - Why can we not avoid making errors in inferential statistics?
  - Why do we need the assumption to be met?
  - How to conduct a test?
  - What is the difference between parameter, estimator, and estimate?
  - What is the role of a model?

  If you feel you are a bit shaky regarding these concepts, find the review notes in STAT 252 on my website for a comprehensive review.

- In this course we add to your statistical "toolbox".
  - How are estimators chosen?
  - More probability theory
  - More models
  - Analysis of categorical data (descriptive and inference)
  - Using SPSS in the process

What are your expectations?

1.1 Models

In statistics we use models to describe the population and then we use sample data, to estimate parameters in the model, and conduct tests relating to these parameters.
The models are usually characterized by the response variable (the variable we want to learn about), and the independent(explanatory) variables (factors), which might or might not be related to the response variable.
All models considered in STAT 252 and STAT 266 assumed that we had one numerical response variable, and either categorical factors (ANOVA) or numerical (and categorical) predictors (Linear Regression).

In this course we will discuss models for categorical response variables.
In a course on multivariate statistics we would discuss models with more than one response variable.

1.2 Categorical Data

Categorical variables fall within two categories: nominal and ordinal

- A variable is called ordinal, if its scale reflects an order.
  E.g: attitude towards quality of education (excellent, good, fair, poor)
  age (child, youth, young adult, middle aged, old adult)
  length of vacation (too short, short, just right, long, too long)
  Likert scale items (strongly agree, agree, neutral, disagree, strongly disagree)
  Did you take Stat 252? (yes,no)

  The statistical tools used for this type of variable preferably takes this ordering into account.

- A variable is called nominal, if such an order does not exist.
  E.g.: religious affiliation (Catholic, Jewish, Protestant, Muslim, none, other)
  preferred grocery store (Safeway, Superstore, Save-on-foods, Greenwood, other)

  The order of the listing is irrelevant, and only statistical tools should be used which do not depend on the order.

Tools which can be used for nominal data can also be used for ordinal variables (even though some information is lost), but models which are designed for ordinal data are not appropriate for nominal data.

1.3 Probability Distributions for Categorical Variables

In this course we study tools to analyze categorical data, before we can go into statistics the models for such populations have to be developed. Models for the population are always claims about the distribution of the variable in the population. Therefore we need to study the distributions for describing categorical variables.

Remember distributions describe (categorical) random variables by

1. indicating the possible values, and
2. providing the probabilities for the values to occur.

1.3.1 Binomial Distribution

- trial ends either in success or failure (two possible outcomes)
- in each trial the probability for success is π
  (here π is not the number 3.1415..., but the probability parameter of the binomial distribution)
- repeat the trial n times independently from each other
- Y is the random variable that gives the number of successes

Then Y has a binomial distribution with possible outcomes $\{0, 1, 2, \ldots, n\}$ and for $y \in \{0, 1, 2, \ldots, n\}$

$$P(Y = y) = p(y) = \binom{n}{y} \pi^y (1 - \pi)^{n-y}$$

with

$$\binom{n}{y} = \frac{n!}{y!(n-y)!}, \quad n! = n(n-1)(n-2)\ldots 1$$

The mean number of successes is

$$\mu = n\ \pi$$

and the standard deviation is

$$\sigma = \sqrt{n\pi(1-\pi)}$$

Example 1
Assume we are looking for a model to describe treatment success (for example for bypass surgery). Assume the success rate is 0.8.
Now we look at 5 patients (each patient represents a trial), the probability for success for each patient is $\pi = 0.8$, and $n = 5$. Y is the number of successes in the 5 surgeries.
The probability for no success in the five surgeries is:

$$P(Y = 0) = p(0) = \binom{5}{0} 0.8^0 (1 - 0.8)^5 = \frac{5!}{0!5!} 1 \cdot 0.2^5 = 0.2^5 = 0.00032$$

The probability for two success in five surgeries is:

$$p(2) = \binom{5}{2} 0.8^2 (1 - 0.8)^3 = \frac{5!}{2!3!}\ 0.64 \cdot 0.008 = \frac{5 \cdot 4}{2}\ 0.00512 = 0.0512$$

How many successes would we expect in average in 5 surgeries? (this is the mean) $\mu = n \cdot p = 5(0.8) = 4$
If we would look at the number of successful surgeries in 5 patients again and again how much spread would we see in the number of successes? $\sigma = \sqrt{5 \cdot 0.8 \cdot 0.2} = 0.8944$
(if you look at 20 patients it would be larger – 1.789)

1.3.2 Multinomial Distribution

Often there are more than two possible outcomes:

- trial could end in one of c different outcomes ($c = 2$ is binomial)
- in each trial the probability for outcome i is π_i, $\sum_i \pi_i = 1$
- repeat the trial n times independently from each other
- Y_i is the random variable that gives the number of observations in category i

Then the tuple $(Y_1, Y_2, \ldots, Y_c)$ has a multinomial distribution with the probability that n_1 fall into category 1, n_2 fall into category 2, ..., n_c fall into category c, with $\sum_i n_i = n$, is then

$$P(Y_1 = n_1, Y_2 = n_2, \ldots, Y_c = n_c) = p(n_1, n_2, \ldots, n_c) = \frac{n!}{n_1! n_2! \ldots n_c!} \pi_1^{n_1} \pi_2^{n_2} \ldots \pi_c^{n_c}$$

The mean number of observations in category i is then

$$\mu_i = n \ \pi_i$$

and the standard deviation is

$$\sigma_i = \sqrt{n\pi_i(1 - \pi_i)}$$

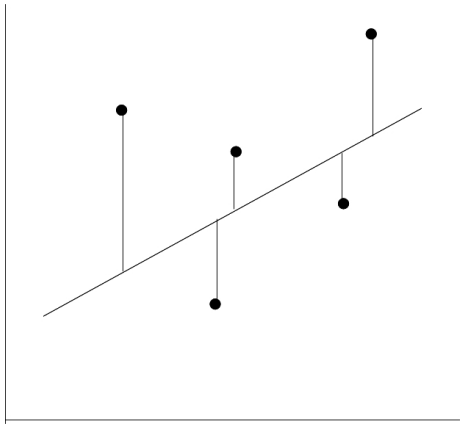
Example 2
Instead of just calling a surgery a success or failure, we add a category partial success, therefore $c = 3$. For the three categories, assume we know $\pi_f = 0.2, \pi_s = 0.5, \pi_{ps} = 0.3$ (total is one).
Then the probability for one failure, one partial success, and 3 successes in 5 surgeries is

$$P(Y_f = 1, Y_{ps} = 1, Y_s = 3) = p(1, 1, 3) = \frac{5!}{1!1!3!} \ 0.2^1 0.3^1 0.5^3 = 20 \cdot 0.2 \cdot 0.3 \cdot 0.125 = 0.15$$

In five surgeries we expect in average $5(0.2) = 1$ failure, $5(0.3) = 1.5$ partial successes, and $5(0.5) = 2.5$ successes.

1.4 Principle of Maximum Likelihood

In practice the probabilities for the different categories are unknown, and one task in inferential statistics is the estimation of these probabilities from sample data.
The question arises how should one use the sample data best to find such estimates.
In linear regression the least squares estimator is introduced. Choosing estimates, such that the estimated model has the "smallest distance" to the data.



In categorical data analysis another strategy for choosing an estimator called Maximum Likelihood is introduced.
This method will be demonstrated by finding an estimator for the probability, π, of a binomial distribution.
Assume the data collected consist of 10 trials including 7 observed successes. What would be a good estimate for the probability of a success in a single trial?

- If it would be true that $\pi = .2$, the probability for observing 7 successes would be

$$P(7) = \binom{10}{7} 0.2^7 (1 - 0.2)^3 = 0.00079$$

- If $\pi = .5$, the probability for observing 7 successes would be

$$P(7) = \binom{10}{7} 0.5^7 (1 - 0.5)^3 = 0.11719$$

- If $\pi = .7$, the probability for observing 7 successes would be

$$P(7) = \binom{10}{7} 0.7^7 (1 - 0.7)^3 = 0.26683$$

For the different choices of values for π, $P(7)$ is largest for $\pi = .7$. This is not only true for the chosen values, but for all possible choices for π (namely $\pi \in [0, 1]$).
This is telling us that the probability for observing 7 successes in 10 trials is the most likely outcome in the case when $\pi = 0.7$. Therefore a reasonable estimate for π, would be $\hat{\pi} = 0.7$.

The Maximum Likelihood Estimate is the value of the parameter that results in the highest likelihood for the observed data to occur.

More general: The **likelihood function** $L(\pi)$ gives the probability of observing the sample data depending on the value of the population parameter.

The **Maximum Likelihood estimate** is the value $\hat{\pi}$ which makes the likelihood function as large as possible, or the parameter value for which the probability of the observed outcome takes its largest value.

The **Maximum Likelihood estimator** is then the function, which assigns to sample data the Maximum Likelihood estimate.

Example 3
Find an estimate for π, the proportion of flights being on time during the Thanksgiving weekend, based on a random sample of 300 flights within Canada, of which 170 were on time.

Because the random variable describing if a single flight is on time has two possible outcomes (yes,no), and the outcomes for different flights can be considered independent, the distribution of the **number** of flights being on time follows a binomial distribution.
Once n= number of flights observed and y= number of flights on time are known, the likelihood function is according to the binomial distribution:

$$L(\pi) = \binom{n}{y}\pi^y(1-\pi)^{n-y} \underbrace{=}_{\text{here}} \binom{300}{170}\pi^{170}(1-\pi)^{130}$$

The function is largest when $\pi = \hat{\pi} = y/n$ (if you took calculus, confirm this result). Therefore the ML estimator for π is

$$MLE(\pi) = \hat{\pi} = p = \frac{y}{n}$$

And the ML estimate for π from our example is

$$p = \hat{\pi} = \frac{170}{300} = 0.5666$$

The ML estimator is a random variable (like the z-score, t-score, or sample mean). Random variables are in general described by their (sampling) distribution.
Distributions tell us the possible values of a random variable, and how likely these are to occur.

Estimators based on the ML method are popular because of their properties for large samples:

- They are consistent, when increasing the sample size they "converge in probability" to the true value of the parameter.
- Under all consistent estimators they are most efficient, they have the smallest standard errors (= standard deviation of the estimator).
- They are approximately normally distributed.

1.5 Confidence Interval for π

Remember the purpose of calculating confidence intervals? Let $z_{\alpha/2}$ denote the value having right-tail probability of $\alpha/2$ for a standard normal distribution.
The standard error of $\hat{\pi}$ is given by

$$SE(\hat{\pi}) = \sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}$$

Then a $(1-\alpha)\times 100\%$ confidence interval for π is given by

$$\hat{\pi} \pm z_{\alpha/2} SE(\hat{\pi})$$

What does it mean to be x% confident?

This confidence interval is based on the fact that

$$Z = \frac{\hat{\pi} - \pi}{SE(\hat{\pi})}$$

is approximately normally distributed, which is of course not true for small sample sizes. Therefore a different formula for finding confidence intervals for small samples should be used.

For small samples the **"plus four method"** (or Agresti-Coull) confidence interval gives better results.
Here we pretend we have four more observations of which 2 are successes and 2 are failures, then we use for the resulting numbers the formula above. Simulations have shown that for small samples, the success rate

$$\frac{\text{number of confidence intervals including the true value of } \pi}{\text{number of confidence intervals calculated}}$$

of this method is closer to $1-\alpha$ then for the original formula.

Example 4
To estimate the proportion, π, of businesses that provide incentives to carpool to work, students polled 15 companies, of these 2 have programs in places to encourage car pooling.
Then $\hat{\pi} = 2/15 = 0.133$ is the ML estimate for π, and $SE(\hat{\pi}) = \sqrt{0.133(1-0.133)/15} = 0.08768$. For a 95% confidence interval, $\alpha = 0.05$, $\alpha/2 = 0.025$, and $z_{0.025} = 1.96$. The 95% confidence interval is then

$$0.133 \pm 1.96 \cdot 0.08768 \leftrightarrow 0.133 \pm 0.1718 \leftrightarrow [-0.0388, 0.3048]$$

This does not make sense, since the true proportion can not be negative!

The plus four method delivers the following interval $\tilde{\pi} = 4/19 = 0.210$, $SE(\tilde{\pi}) = 0.0936$

$$0.210 \pm 1.96 \cdot 0.0936 \leftrightarrow 0.210 \pm 0.184 \leftrightarrow [0.027, 0.395]$$

That is better! We are 95% confident that the proportion of businesses which actively support car pooling falls between 0.027 and 0.395, i.e. between 2.7% and 39.5%.

1.6 Test for π

Remember the purpose of a statistical test?
Remember the 6 steps in a complete statistical test?

A Large Sample Test concerning a Proportion π

1. **Hypotheses:**

Test type	
Upper tail	$H_0 : \pi \leq p_0$ versus $H_a : \pi > p_0$
Lower tail	$H_0 : \pi \geq p_0$ versus $H_a : \pi < p_0$
Two tail	$H_0 : \pi = p_0$ versus $H_a : \pi \neq p_0$

   Choose α.

2. **Assumption:**Random sample and, the sample size n is large, that is that $n\hat{\pi} > 5$ and $n(1-\hat{\pi}) > 5$.

3. **Test statistic:** Let p_0 be a value between zero and one and define the test statistic

$$z_0 = \frac{\hat{\pi} - p_0}{\sqrt{(p_0(1-p_0))/n}}$$

4. **P-value:**

Test type	**p-value**
Upper tail	$P(z > z_0)$
Lower tail	$P(z < z_0)$
Two tail	$2 \cdot P(z > abs(z_0))$

5. **Decision:**

   If P-value$\leq \alpha$ then reject H_0
   If P-value$> \alpha$ then do not reject H_0

6. **Context.**

Example:
Do the data provide sufficient evidence that less than 20% of businesses provide incentives for car-pooling?

1. We want to test (putting what we are asking in the alternative hypothesis H_a)
   $H_0 : \pi \geq 0.2$ vs. $H_a : \pi < 0.2$ at a significance level of $\alpha = 0.05$.

2. The sample size is $n = 15$ and $y = 2$, so that $\hat{\pi} = 0.133$, $n\hat{\pi} = 2$ and $n(1 - \hat{\pi}) = 13 > 5$, so the assumptions are not met (assuming the sample was randomly chosen).

   Why is this problematic? Now the distribution of the z-score is not approximately normal and the p-value will be wrong, therefore should we be able to reject H_0, the error probability might be larger than α or smaller, we do not know.

   We should use the plus four method with $\tilde{\pi} = 4/19 = 0.210$, $SE(\tilde{\pi}) = 0.0936$

3. Then
$$z_0 = \frac{0.210 - 0.2}{0.0936} = 0.106$$

4. Now calculate the P-value, according to the choice of H_a it is a lower tail test, so the P-value is the lower tail probability. $\mathrm{P} - \mathrm{value} = P(z < 0.106) = 0.5438$ (from table normal distribution table.)

5. Decision: Since 0.54388=P-value$> 0.05 = \alpha$, H_0 is not rejected.

6. Context: At significance level of 5% the sample data do not provide sufficient evidence that less than 20% of businesses provide incentives for car pooling.

   Remember when not rejecting the null hypothesis we do not have any information on the error probability, therefore the careful wording (not committing to the decision).

1.7 Wald, Likelihood Ratio, and Score Inference

Generalization for arbitrary parameters.

Three different types of test statistics/confidence intervals are introduced. These can potentially lead to three different answers for the same question, which can be quite confusing. Of the three answers, which is correct?
I compare this to having 2 clocks: with one clock one "knows", what time it is, but with two clocks (not showing the same time) one is left in doubt.
In some scenarios, it won't be possible to calculate all three options.

1. **Wald:** Let β be an arbitrary parameter, and $\hat{\beta}$ the ML-estimator for β (remember $\hat{\beta}$ is a random variable). The Wald score is based on the SE for the ML-estimate.

   Then
   $$Z = \frac{\hat{\beta} - \beta_0}{SE(\hat{\beta})}$$
   is approximately normally distributed (because $\hat{\beta}$ is normal, because it is an ML-estimator).

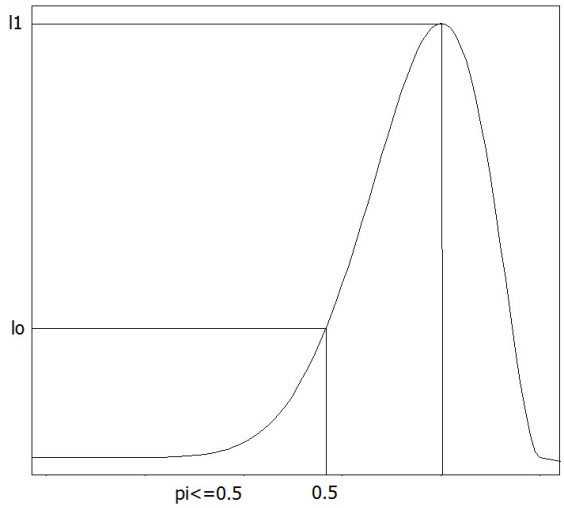
   We also know that in this case Z^2 is approximately χ^2 distributed with $df = 1$. This type of statistic is called a Wald statistic, and tests based on Z or Z^2 are called Wald tests.

2. **Likelihood Ratio:** An alternative test is based on the likelihood function.

   - Let l_0 be the maximum of the likelihood function for parameter values from H_0.
   - Let l_1 be the likelihood function for $\hat{\beta}$, that is the largest possible value for the likelihood function.

   The l_1 is always at least as large as l_0.

   The likelihood ratio test statistics is then
   $$-2\log(l_0/l_1)$$



If H_0 is not true, we should expect that l_0 is much smaller than l_1, then the ratio l_0/l_1 is much smaller than 1, therefore the log (meaning the natural logarithm: ln) is much smaller than 0, therefore the likelihood ratio test statistics will be very large (and positive).

This test statistic can be proved to be approximately χ^2 (with $df = 1$) distributed when H_0 is true (cool!)

3. **Score Inference**

   The score inference statistic is based on the standard error of the ML-estimator, if H_0 would be true, therefore using the standard error, for H_0.

   For example: When testing $H_0 : \pi = 0.5$, then $SE = \sqrt{0.5(1-0.5)/n}$.

   Therefore the test introduced above is a score test.

Comments:

- For the linear regression model with a normal error Wald, Likelihood Ratio and Score tests are all the same.
- The calculated P-values are usually only approximations of the true P-values, because we use approximative distributions. In general the larger the sample sizes the better the approximation. Therefore only report at most 3 decimal places.
- Each method can be used to find confidence intervals for β:

  A $(1-\alpha)\times$100% confidence interval is the set of all β_0, so that $H_0 : \beta = \beta_0$ can not be rejected at significance level of α. How does that make sense?

Example 5
We are using a new sample to test if the proportion of businesses which give incentives for carpooling is different from 0.2. The larger sample includes $n = 100$ businesses and $y = 15$. We will use all three tests for testing

- $H_0 : \pi = 0.2$ versus $H_a : \pi \neq 0.2$, at $\alpha = 0.05$.
- Assumptions: Random sample! Relatively large sample size. 100 is usually large enough
- Test statistic:

  1. Wald test: $\hat{\pi} = 15/100 = 0.15$, $SE(\hat{\pi}) = \sqrt{0.15(1-0.15)/100} = 0.0357$

     $$z_0 = \frac{0.15 - 0.2}{0.0357} = -1.400$$

  2. Likelihood ratio: As discussed in section 1.4 the likelihood function is

     $$l(\pi) = p(15) = \binom{100}{15}\pi^{15}(1-\pi)^{85}$$

     Then for π_0 from H_0

     $$l_0 = l(0.2) = 0.04806$$

and for $\hat{\pi}$

$$l_1 = l(0.15) = 0.11109$$

therefore

$$\chi_0^2 = -2\ln(l_0/l_1) = -2\ln(0.43262) = 1.6758, \quad df = 1$$

3. Score test: $SE(\pi_0) = \sqrt{0.2(1-0.2)/100} = 0.04$ (using 0.2 from H_0)

$$z_0 = \frac{0.15 - 0.2}{0.04} = -1.25$$

4. P-value:
   1. Wald: $p = 2 \times P(Z < -1.4) = 2(0.0808) = 0.1616$ (normal distribution table)
   2. likelihood ratio: $P = P(\chi^2 > 1.6758) > 0.10$ (chi-squared table)
   3. score test: $P = P(Z < -1.25) = 2(0.1056) = 0.2112$ (normal distribution table)
5. Decision: With $\alpha = 0.05$ Do not reject H_0 with any of the three tests.
6. At significance level of 5% the data do not provide sufficient evidence that the proportion of businesses that support carpooling is different from 0.2.

Would 0.2 be in the 95% Wald confidence interval for π? Yes, because the Wald test for $H_0 : \pi = 0.2$ did not result in the rejection of H_0 at significance level of $\alpha = 5\%$.