STAT231 Notes

Richard Wu

1 Miscellaneous

1.1 Types of Variates

Variates can be separated into categorical (non-numerical) and numerical. Numerical is further divided into discrete (integer values) and continuous. Note encoded categorical variables are still considered categorical (and not discrete). Categorical can also be ordinal (with order) or non-ordinal.

1.2 Unbiased Estimate of σ^2

The unbiased estimate of σ^2 is actually s^2 or s_e^2 (for SLRM).

2 May 8, 2017

2.1 Calculating percentiles

Let

$$m = (n+1) \times p$$

m	the index for the p percentile
n	sample size
p	desired percentile

If m is integer, take the mth element. Otherwise, take the average of the $\lfloor m \rfloor$th and $\lceil m \rceil$th element.

3 May 10, 2017

3.1 Variance

$$
\begin{aligned}
s^2 &= \frac{1}{n-1}\sum^n (y_i - \bar{y})^2 \\
&= \frac{1}{n-1}\sum^n (y_i^2 - 2y_i\bar{y} + \bar{y}^2) \\
&= \frac{1}{n-1}(\sum^n y_i^2 - 2\bar{y}\sum^n y_i + n\bar{y}^2) \\
&= \frac{1}{n-1}(\sum^n y_i^2 - 2\bar{y}(n\bar{y}) + n\bar{y}^2) \\
&= \frac{1}{n-1}(\sum^n y_i^2 - 2n\bar{y}^2 + n\bar{y}^2) \\
&= \frac{1}{n-1}(\sum^n y_i^2 - n\bar{y}^2)
\end{aligned}
$$

3.2 Affine transformation

Original data: $\{x_1, x_2, \ldots, x_n\}$

$$y_i = a + bx_i \qquad \forall i = 1, \ldots, n$$

Our statistical values change:

$$
\begin{aligned}
\bar{y} &= a + b\bar{x} \\
s_y^2 &= b^2 s_x^2 \\
s_y &= |b|x \\
Range &= (a + bx_{max}) - (a + bx_{min}) \\
&= b(x_{max} - x_{min})
\end{aligned}
$$

3.3 Measure of Symmetry

Skewness

- 0 is perfectly symmetric ($mean \approx median$)
- +'ve has long right tail ($mean > median$)
- -'ve has long left tail ($mean < median$)

3.4 Measure of Kurtosis

Kurtosis measures *frequency of extreme observations* compared to the Gaussian distribution.

For a Gaussian distribution, 99% of observations lie within $\bar{y} \pm 3s$.

Kurtosis for perfect Gaussian **is always** 3. If $K >> 3$, then frequency of extreme observations are more than Gaussian.

3.5 Measures of Association

Objective to find the strength of association between X and Y.

Categorical

	Positive	Negative
Smoker	y_{11}	y_{12}
Non-Smoker	y_{21}	y_{22}

Relative Risk

$$RR = \frac{\frac{y_{11}}{y_{11}+y_{12}}}{\frac{y_{21}}{y_{21}+y_{22}}}$$

For high and low values of RR, there is evidence of association.

4 May 12, 2017

4.1 Sample Correlation Coefficient

To calculate r_{xy} the sample correlation coefficient

$$\frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{[\sum(x_i - \bar{x})^2]^{\frac{1}{2}}[\sum(y_i - \bar{y})^2]^{\frac{1}{2}}}$$

which denotes the **direction and strength** of the *linear* association between X and Y.

Note when $x_i > \bar{x}$ and $y_i < \bar{y}$, we have

$$(x_i - \bar{x})(y_i - \bar{y}) < 0$$

and similarly for $x_i < \bar{x}$ and $y_i > \bar{y}$. This means the **numerator denotes the direction** of the relationship.

Denominator guarantees r_{xy} is between 1 and -1, inclusively. More formally

$$-1 \leq r_{xy} \leq 1$$

and for a given $y_i = a + bx_i \forall i$ relationship

$$r_{xy} = \begin{cases} 1 & b > 0 \\ -1 & b < 0 \end{cases}$$

5 May 15, 2017

5.1 CDFs

You may calculate the **percentile** by simply taking $F(percentile)$. The **mode** can be inferred by the largest "jump" in a discrete CDF.

5.2 Box Plots

Use box plots to compare distribution shape of two or more data sets side-by-side.

Whiskers mark *min* and *max* data points. The box itself begins and ends at $Q1$ and $Q3$, respectively, with a line at the median or $Q2$.

5.3 Scatter Plots

Maps bivariate distribution (x and y) to discern whether there is a relationship or not.

6 May 19, 2017

6.1 θ

θ is a given attribute of the population that we want to find out. We can never find out attribute unless we sample the entire population.

Objective: To find an "estimate" of θ based on $\{y_1, \ldots, y_n\}$.

6.2 Maxmimum Likelihood Estimate (MLE)

Question: What is the "most likely" values of θ, given your sample?

$\hat{\theta}$ is the **maximum likelihood estimate or MLE**. That is $\hat{\theta}(y_1, \ldots, y_n)$ is a known value if given the sample.

Step 1 is always setting up a statistical model to estimate the likelihood function. **Model** is the "identification" of the random variable Y_i from which y_i is an outcome.

Moreover, θ is also a parameter of that random variable.

$$Y_i \sim f(y_i; \theta) \qquad i = 1, \ldots, n$$

where f is the distribution function of Y_i.

Example. A coin is tossed. Let θ be the *probability of getting a HEAD.*

$$\theta = \left\{ \begin{array}{cc} & \frac{1}{3} \\ \text{or} & \frac{2}{3} \end{array} \right\}$$

these are the only two (theoretical) possilibities. Note we do not know what θ is yet.

The coin is tossed 200 times, and we observe $y = 140$ heads in the *sample*.
$\hat{\theta} = \frac{2}{3} = MLE$ is the "most likely" value of θ **given our sample**.

6.3 Likelihood Function

Definition. The **likelihood function** $L(\theta; y_1, \ldots, y_n)$ is the probability of observing the sample as a function of θ.

Example. Continuing from the previous example:
Suppose $\theta = \frac{1}{3}$. What is the chance of observing our sample?

$$\binom{200}{140}\theta^{140}(1-\theta)^{60} = \binom{200}{140}\frac{1}{3}^{140}(\frac{2}{3})^{60}$$

If $\theta = \frac{2}{3}$, similarly

$$\binom{200}{140}\frac{2}{3}^{140}(\frac{1}{3})^{60}$$

Note the probability for our sample for when $\theta = \frac{2}{3}$ is greater than that for when $\theta = \frac{1}{3}$. So $\theta = \frac{2}{3}$ is our MLE.

Example. Sample people to find out who likes Trump. So θ is the probability someone likes Trump.
Given this sample

$$\{N, N, N, N, Y, Y, N, N, Y\}$$

what is the estimate for θ?
Note that our likelihood function for a given θ is

$$L(\theta) = (1-\theta)^7\theta^3$$

we need to maximize this with respect to θ.

6.4 Log-Likelihood Function

Introduce the **log-likelihood function** which is

$$l(\theta) = log(L(\theta))$$

all logs are *base e*.

Example. So the $l(\theta)$ for our Trump sample is

$$l(\theta) = 7ln(1-\theta) + 3ln(\theta)$$

6.5 First Order Condition

To maximize $L(\theta)$, we can take the $\frac{dl}{d\theta} = 0$ to solve for $\hat{\theta}$

Example.

$$\frac{-7}{1-\theta} + \frac{3}{\theta} = 0$$
$$\frac{7}{1-\theta} = \frac{3}{\theta}$$
$$\theta = 0.3$$

So our MLE $\hat{\theta} = 0.3$, which makes sense since there're 3 Ys.

In a binomial distribution, *sample proportion is the MLE for the population proportion.*

6.6 Poisson Distribution Example

Example. Let θ be the number of accidents at an intersection during rush hour in the month of May. θ is the same as λ in $Pois(\lambda)$.

Let our model by a poisson distribution $Y\ Pois(\theta)$.

What is our $L(\theta)$? Given the sample

$$\{2, 0, 1, 0, 3\}$$

from the PDF for the poisson distribution, we get for all 5 elements

$$\frac{e^{-\theta}\theta^2}{2!} \cdot \frac{e^{-\theta}\theta^0}{0!} \cdot \ldots \cdot \frac{e^{-\theta}\theta^3}{3!}$$

which results in the likelihood function

$$L(\theta) = \frac{e^{-5\theta}\theta^6}{2!0!1!0!3!}$$

The log-likelihood function is

$$l(\theta) = -5\theta + 6ln(\theta) - ln(2!0!1!0!3!)$$

which when we take $\frac{dl}{d\theta} = 0 \rightarrow -5 + \frac{6}{\theta} = 0$, we get

$$\hat{\theta} = \frac{6}{5}$$

For the poisson distribution, the **sample mean is the MLE** of θ.

Example. Jeopardy problem. Sample is the number of games a Canadian participated in. You must win a game to play another game. Let θ be the probability that a Canadian wins Jeopardy. Our sample is the number of wins of a sample of Canadians

$$\{2, 1, 1, 3\}$$

This maps to the probabilities

$$\theta(1-\theta) \cdot (1-\theta) \cdot (1-\theta) \cdot \theta^2(1-\theta)$$

7 May 23, 2017

7.1 (Discrete) General Likelihood Function

The likelihood function (assuming **independence** and **discrete**) can be generalized in terms of each measurement in the sample. From our initial definition of the likelihood function

$$\begin{aligned} L\theta; y_1, \dots, y_n) &= P(Y_1 = y_1, Y_2 = y_2, \dots, Y_n = y_n) && \text{definition} \\ &= P(Y_1 = y_1) \cdot P(Y_2 = y_2) \cdot \dots \cdot P(Y_n = y_n) && \text{independent} \\ &= f(y_1) f(y_2) \dots f(y_n) && \text{notation} \end{aligned}$$

So we have the general equation

$$L\theta; y_1, \dots, y_n) = \prod_{i=1}^{n} f(y_i; \theta)$$

which is the *product of the distribution functions* evaluated at the sample points ($Y_i = y_i$).

Definition. $\hat{\theta}$ is called the MLE (Maximum Likelihood Estimate) if $\hat{\theta}$ **maximizes** $L(\theta)$.

7.2 (Discrete) *General* Poisson Example

Example. Data is drawn from a Poisson distribution with unknown mean μ. The data set is $\{y_1, \dots, y_n\}$ is independently drawn. Based on your sample, what is $\hat{\mu}$.

Note

$$P(Y = y) = \frac{e^{-\mu}\mu^y}{y!} \qquad y = 0, 1, 2, \dots$$

So we have the likelihood function

$$L(\theta; y_1, \dots, y_n) = \frac{e^{-\mu}\mu^{y_1}}{y_1!} \cdot \frac{e^{-\mu}\mu^{y_2}}{y_2!} \cdot \dots \cdot \frac{e^{-\mu}\mu^{y_n}}{y_n!}$$

Which simplifies to

$$L(\mu) = \frac{e^{-n\mu}\mu^{\sum y_i}}{y_1! y_2! \dots y_n!}$$

where the log-likelihood function is

$$l(\mu) = -n\mu + \sum y_i ln\mu - ln(y_1! y_2! \dots y_n!)$$

Taking the first order condition (FOC) of the log, we have

$$\frac{dl}{d\mu} = 0 \rightarrow -n + \frac{\sum y_i}{\mu} = 0$$
$$\rightarrow \hat{\mu} = \frac{\sum y_i}{n} = \bar{y}$$
$$\rightarrow \hat{u} = \bar{y}$$

For the general *Poisson* problem, $\bar{y}$ **is the MLE for** μ.

7.3 (Discrete) *General* Geometric Example

Example. Let

Y_i # of failures before the 1st success

θ probability of success for each trial

The trials are independent. θ is unknown. Given a sample drawn independent $\{y_1, \ldots, y_n\}$, what is the MLE of θ, that is what is $\hat{\theta}$?

Solution the model for our samples is the geometric distribution $Y_i \sim Geom(\theta)$ for $i = 1, \ldots, n$ i.i.d. So

$$f(y) = P(Y = y) = (1-\theta)^y\theta \qquad y = 0, 1, 2, \ldots$$

Note that when we plug in $Y = \theta$), we get $P(Y = \theta) = \theta$. So we have the likelihood function

$$L(\theta) = (1-\theta)^{y_1}\theta \cdot (1-\theta)^{y_2}\theta \cdot \ldots \cdot (1-\theta)^{y_n}\theta$$
$$= (1-\theta)^{\sum y_i}\theta^n$$

Taking the log or the log-likelihood function

$$l(\theta) = \sum y_i ln(1-\theta) + nln\theta$$

The FOC will be

$$\frac{dl}{d\theta} = 0 \rightarrow \frac{-\sum y_i}{1-\theta} + \frac{n}{\theta} = 0$$

Then solve for θ to find $\hat{\theta}$.

7.4 (Discrete) *General* Binomial Example

Example. Let

$\tilde{n}$ probability of success for each trial

A more concrete example for $\tilde{n}$ is the proportion of left-handed people at UW.

We are doing the experiment n times and we observe y successes.

Based on the sample, what is $\hat{\tilde{n}}$?

This is a *binomial model.*

$$L(\tilde{n}) = \binom{n}{y} \tilde{n}^y (1-\tilde{n})^{n-y}$$

which is the probability of successes in n trials. Taking the log

$$l(\tilde{n}) = ln(\binom{n}{y}) + yln\tilde{n} + (n-y)ln(1-\tilde{n})$$

Then we solve for FOC

$$\begin{aligned} \frac{dl}{d\tilde{n}} = 0 \rightarrow \frac{y}{\tilde{n}} - \frac{n-y}{1-\tilde{n}} = 0 \\ \rightarrow \hat{\tilde{n}} = \frac{y}{n} \end{aligned}$$

which is equivalent to the sample's proportion, which aligns with intuition.

7.5 Some Notes about MLE

Note from the likelihood equation

$$L = P(Y_1 = y_1) \cdot P(Y_2 = y_2) \cdot \ldots \cdot P(Y_n = y_n)$$

The values of L becomes very, very small as n becomes large (composite of many probabilities $\rightarrow$ very small probabilities).

7.6 Relative Likelihood Function

The relative likelihood function is defined as

$$R(\theta) = \frac{L(\theta)}{L(\hat{\theta})}$$

where $\hat{\theta}$ is the MLE.

Note that since $\hat{\theta}$ maximizes L, $R(\theta) \leq 1$. Also, $R(\theta) \geq 0$ for all θ. When $R(\theta) = 1$, then $\theta = \hat{\theta}$.

The relative likelihood function tells us the reasonable values of θ (the values that are close to $\hat{\theta}$.

8 May 24, 2017

8.1 (Continuous) General Likelihood Function

$$L(\theta; y_1, \ldots, y_n) = \prod_{i=1}^{n} f(y_i; \theta)$$

where f is the *density function* of the r.v Y.

8.2 (Continuous) Exponential Example

Example. To find the average (μ) lifespan of a light bulb from a production line.

A sample of n observations are collected $\{y_1, \ldots, y_n\}$. Based on this sample, what is $\hat{\mu}$ (MLE for μ)

Let our model be $Y_i \sim Exp(\mu)$ for $i = 1, \ldots, n$ and Y_is are independent. The density function is

$$f(y) = \frac{1}{\mu} e^{-y/\mu} \qquad \mu > 0, y \geq 0$$

The likelihood function is therefore

$$\begin{aligned} L(\mu) &= \frac{1}{\mu} e^{-y_1/\mu} \cdot \frac{1}{\mu} e^{-y_2/\mu} \cdot \ldots \cdot \frac{1}{\mu} e^{-y_n/\mu} \\ &= \frac{1}{\mu^n} e^{-\frac{1}{\mu} \sum y_i} \end{aligned}$$

Taking the log-likelihood function

$$l(\mu) = -nln(\mu) - \frac{1}{\mu} \sum y_i$$

Taking the first order condition

$$\begin{aligned} \frac{dl}{d\mu} = 0 &\rightarrow -\frac{n}{\mu} + \frac{1}{\mu^2} \sum y_i = 0 \\ &\rightarrow \hat{\mu} = \bar{y} \end{aligned}$$

so for the exponential distribution, the sample mean is the MLE for μ, which aligns with the definition of μ in the exponential density function.

8.3 (Continuous) Gaussian Example

Example. We want to estimate

μ population average IQ of UW profs

σ s.d. of IQ of UW profs

A sample of n profs are selected $\{y_1, \ldots, y_n\}$. Based on your sample, estimate $\hat{\mu}$ and $\hat{\sigma}$.

Assume that the samples are collected independently using a Gaussian model. That is $Y_i \sim G(\mu, \sigma)$. Thus the density function is

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

The likelihood function is

$$\begin{aligned} L(\mu, \sigma) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(y_i-\mu)^2} \\ &= \frac{1}{(2\pi)^{n/2}\sigma^n} e^{-\frac{1}{2\sigma^2}\sum(y_i-\mu)^2} \end{aligned}$$

Thus the log-likelihood function is

$$l = -\frac{n}{2} ln(2\pi) - n ln(\sigma) - \frac{1}{2\sigma^2} \sum (y_i - \mu)^2$$

To maximize l to find $\hat{\mu}$ and $\hat{\sigma}$ we take two derivatives $\frac{\partial l}{\partial \mu} = 0$ and $\frac{\partial l}{\partial \sigma} = 0$. Thus we have

$$\begin{aligned} (1)\ \frac{\partial l}{\partial \mu} &= \frac{1}{\sigma^2} \sum (y_i - \mu) = 0 \\ (2)\ \frac{\partial l}{\partial \sigma} &= -\frac{n}{\mu} + \frac{1}{\sigma^3} \sum (y_i - \bar{y})^2 = 0 \end{aligned}$$

So from (2) we have

$$\begin{aligned} \sum (y_i - \mu) &= 0 \\ \sum y_i - \sum \mu &= 0 \\ \sum y_i - n\mu &= 0 \\ \mu &= \frac{\sum y_i}{n} = \bar{y} \end{aligned}$$

So $\hat{\mu} = \bar{y}$ as expected.

From (1) we have

$$\begin{aligned} -\frac{n}{\sigma} + \frac{1}{\sigma^3} \sum (y_i - \bar{y})^2 &= 0 \\ \hat{\sigma}^2 &= \frac{1}{n} \sum (y_i - \bar{y})^2 \end{aligned}$$

which is the population s.d. equation. Note however that $\sigma^2 \neq s^2$ (denomiators n and $n-1$ are different).

8.4 Properties of MLEs

Invariance Property

If $\hat{\theta}$ is the MLE for θ then $g(\hat{\theta})$ is the MLE for $g(\theta)$ for a continuous g (we will first find $\hat{\theta}$, then plug in $g(\theta)$).

Example. To estimate the 95th percentile of a population (Gaussian), $Y_1, \ldots, Y_n \sim G(\mu, \sigma)$ independent.

Find the MLE for the 95th percentile.

$$
\begin{aligned}
&P(Y \leq A) = 0.95 \\
&P(\frac{Y-\mu}{\sigma} \leq \frac{A-\mu}{\sigma}) = 0.95 \\
&P(z \leq x) = 0.95 \qquad x = \frac{A-\mu}{\sigma}
\end{aligned}
$$

Look at the z score table for the 95th percentile, we get $x = 1.645$. Thus

$$
\begin{aligned}
\frac{A-\mu}{\sigma} &= 1.645 \\
A &= \mu + 1.645\sigma
\end{aligned}
$$

The MLE for A is $\hat{\mu}+1.645\hat{\sigma}$ where $\hat{\mu}$ and $\hat{\sigma}$ are the MLE for μ and σ, respectively (by the invariance property).

Example. Let $Y \sim Bin(200, \pi)$ where $\pi = P(success)$ (unknown).

Let $y =$ number of successes $= 120$ (sample).

Find the MLE for $P(Y > 1)$. Note this is the same as

$$
\begin{aligned}
P(Y > 1) &= 1 - [P(Y = 0) + P(Y = 1)] \\
&= 1 - [\binom{n}{0}(1-\pi)^n + \binom{n}{1}\pi(1-\pi)^{n-1}]
\end{aligned}
$$

9 May 26, 2017

9.1 Uniform Example

Given a uniform distribution $Y \sim Unif(0, \theta)$ where θ is unknown, we take a sample of observations $\{y_1, \ldots, y_n\}$ (independently drawn). What is the MLE of θ?

Note that the density funtion for uniform distributions is

$$
f(y) = \begin{cases} \frac{1}{\theta} & 0 \leq y \leq \theta \\ 0 & \text{otherwise} \end{cases}
$$

The likelihood function is thus

$$
f(y) = \begin{cases} \frac{1}{\theta^n} & 0 \leq y_i \leq \theta, \forall i \\ 0 & \text{otherwise} \end{cases}
$$

this can also be rewritten in terms of $max\{y_1, \ldots, y_n\}$

$$f(y) = \begin{cases} \frac{1}{\theta^n} & \theta \geq max\{y_1, \ldots, y_n\} \\ 0 & \theta < max\{y_1, \ldots, y_n\} \end{cases}$$

Let us denote $\theta^\star = max\{y_1, \ldots, y_n\}$. Note $f(y)$ is simply the first order rational function $\frac{1}{\theta}$ that is bounded between $\theta \in [\theta^\star, \infty)$, and 0 from $\theta \in [0, \theta^\star)$. It is highest at $\theta^\star$. Thus $\hat{\theta} = max\{y_1, \ldots, y_n\}$ is the MLE for θ.

9.2 Model Selection

Given a sample we assume a distribution Y as the model. How do we know that Y is the appropriate model? Use various tests.

Subjective Tests

Numerical Property Tests Compare numerical measures from sample with theoretical properties of the assumed distribution

Example. Given that we assume a Gaussian model, we can check

- Symmetry (mean $\approx$ median; skewness ≈ 0)
- Kurtosis (K ≈ 3)
- Proportion of samples within 2 σ of mean (95% for Gaussian)

Graphical Methods Superimpose theoretical PDF with the sample relative frequency histogram. Similarly for the CDF.

Q-Q plot Bivariate plot of sample quantiles against theoretical quantiles of model

Example. Gaussian model: We need only compare the sample quantiles with theoretical quantiles of $Z = G(0, 1)$.

We want to plot (z_α, y_α) where z_α is the α^{th} quantile of the Z and y_α is the α^{th} quantile of the sample.

Note if our sample is Gaussian, it will form a **straight line** (not necessarily $y = x$ since the model is just the standard Gaussian model). Note this checks if the quantiles are linear functions of each other.

Observed vs Expected Frequency

10 May 29, 2017

Examples

(1) Suppose we are interested in attitude of recent (after 2010) immigrants in Canada towards cops. Sample of 100 immigrants drawn from KW area and survey is conducted.

(2) Find out whether there is an association betwen smoking habits of parents with that of teenage sons/daughters.

(3) Predict student i's STAT231 final score based on past performance of a sample of students with similar GPA.

We will use **PPDAC**.

10.1 Problem

The **target population** is the population whose variate we are interested in.
In example (1), it is all Canadian immigrants since 2010.
There are three types of problems

Descriptive Where we want to estimate the unknown parameter of the population (example (1)).

Causative Where we are trying to find the presence (or absence) of association between two variables X and Y (example (2)).

Predictive We are trying to predict a value of your r.v. (typically) based on past observation (example (3)).

There are two types of variates

Response Variate The variables whose variability we are trying to explain (dependent variable)

Explanatory Variate The variable that is used to explain the response (independent variable)

So the (response) *variate* in example (1) is whether the immigrant has a positive/negative attitude towards cops.
Attributes in example (1) are properties of immigrants with a positive attitude.

10.2 Plan

Questions to ask

1. What is the study population?
2. What is our sampling protocol? (How do I collect my data?)

The **Study population** is the population from with your sample is drawn.
In example (1), the *target population* is all Canadian immigrants since 2010, whereas the *study population* is all Canadian immigrants since 2010 who live in the KW area. In most cases, the study population *is a subset* of the target population.
An exception is a medical test for a drug. The *target* population is human beings with a high heart rate, but the *study* population is mice in a laboratory.
How is the data collected?

Experimental plan the data collector controls some variables (physical sciences)

Observational plan the data collector has no control over the variables (social sciences)

Random sampling each member of the *study population* has an equal chance of being chosen.

10.3 Data

Types of data we have collected: binary, numerical, discrete, continuous.

We must ensure our data is **unbiased**. Pay special attention to *outliers* and *missing* observations.

Systematic errors are called *bias*. Avoid systematic errors!

10.4 Analysis

Estimation, hypothesis testing and predictions for each type of problem, respectively.

Study error: difference in value between the attribute of the *study* population and the *target* population. **Sampling error:** difference in the value between *study* and the *sample*.

Our goal is to analyze and minimize these errors.

10.5 Conclusion

We have to write our conclusions which is understandable for non-statisticians. That comes with useful graphs.

11 May 30, 2017

11.1 Likelihood Intervals

We have an unknown parameter θ of interest. Sample $\{y_1, \ldots, y_n\}$ drawn independently. Assume a chosen model $Y_i \sim f(y_i; \theta)$ is correct.

Objective What are the reasonable values of θ based on my sample? To construct an interval $[l, u]$ such that the unknown parameter θ would lie in the interval with a high degree of confidence. Note that l and u are functions of $\{y_1, \ldots, y_n\}$.

Method 1 Using the likelihood function

What are the values of θ that are "close" to $\hat{\theta}$? Any values of θ such that $L(\theta)$ is close to $L(\hat{\theta})$, we will consider that θ to be "reasonable".

Definition. Take any $p \in (0, 1)$. A $100p\%$ likelihood interval is

$$\{\theta : R(\theta) \geq p\}$$

where $R(\theta)$ is the *relative likelihood function*.

Example. Given $p = 0.5$ (50% likelihood interval), we have to find all θs such that

$$R(\theta) \geq 0.5$$
$$\frac{L(\theta)}{L(\hat{\theta})} \geq 0.5$$
$$L(\theta) \geq 0.5L(\hat{\theta})$$

The value of the likelihood function at θ is at least half the value of L at $\hat{\theta}$.

Example. Binomial problem for θ (proportion of success).
Let R_1 represent $n_1 = 100$ and $y_1 = 40$ successes.
Let R_2 represent $n_2 = 1000$ and $y_2 = 400$.
Note $\hat{\theta} = 0.4$ so θ is maximized for both (parabola with peak at $\theta = 0.4$.
Note that the shape of R_2 will be *thinner* than the shape of R_1 since we have more observations and thus the interval intuitively should be smaller.

11.2 Conventions for Likelihood Intervals

We classify the following $R(\theta)$ as such

$R(\theta) \geq 0.5$ θ very plausible

$R(\theta) \in [0.1, 0.5)$ θ plausible

$R(\theta) \in [0.01, 0.1)$ θ implausible

$R(\theta) < 0.01$ θ very implausible

Anything $\geq 10\%$ is plausible!

Example. $Y \sim Bin(n, \theta)$, $n = 500$, $y = 200$, is $\theta = 0.5$ plausible?

Step 1 Construct the likelihood function $L(\theta) = \binom{500}{200}\theta^{200}(1-\theta)^{300}$

Step 2 Calculate the MLE using Step 1

Step 3 Calculate $R(\theta)$

$$\begin{aligned} R(\theta) &= \frac{L(\theta)}{L(\hat{\theta})} \\ &= \frac{\binom{500}{200}\theta^{200}(1-\theta)^{300}}{\binom{500}{200}0.4^{200}(0.6)^{300}} \end{aligned}$$

If the $R(\theta)$ is bi-modal (two local maxima), then the likelihood interval (LI) is a union of two disjoint intervals e.g. $[l_1, u_1] \cup [l_2, u_2]$.

12 June 2, 2017

12.1 Coverage and Confidence Intervals

Method 2 Sampling distributions

Previously with likelihood intervals a interval where θ lies in with a high degree of confidence (based on the $R(\theta) \geq p$ for some $p \in (0, 1)$).

Now we are given a pre-specified probability (e.g. 90%, 95%) and we want to estimate the random variables L and U such that

$$P(L \leq \theta \leq U) = 0.95$$

L and U are estimated using our sample (l, u).

Example. We want to construct a 95% CI (confidence interval) for μ (population mean) of the average score in STAT231.

A sample of 36 students are drawn independently $\{y_1, \ldots, y_36\}$.

Based on the data set, what is the confidence interval?

Our model can be Gaussian

$$Y_i \sim G(\mu, \sigma), i \in \{1, \ldots, 36\}$$

Assumption: $\sigma = 7$ is known (population standard deviation). Thus we have

$$Y_i \sim G(\mu, 7^2)$$

where $\hat{\mu} = MLE = \bar{y}$. Note that we can denote the r.v. $\bar{Y}$ for which $\bar{y}$ (the number calculated from each sample) is an outcome.

Say we wanted to calculate θ (in Binomial model), and we had $n = 500$ and $y = 220$ thus $\hat{\theta} = 220/500 = 0.44$. Then θ is the unknown population proportion, $\hat{\theta}$ is the MLE (# from sample), $\tilde{\theta}$ is an r.v. from which $\hat{\theta}$ is an outcome.

We call $\bar{Y}$ the **estimator** and $\bar{y}$ an **estimate**. the MLE is also an estimate.

If $Y_1, \ldots, Y_n \sim G(\mu, \sigma)$ independent, then the mean of the distributions is

$$\bar{Y} = \frac{1}{n} \sum_{i=0}^{n} Y_i$$

$$\bar{Y} = G(\mu, \frac{\sigma}{\sqrt{n}})$$

Example. So from the previous example, $\bar{Y} \sim G(\mu, \frac{7}{6})$.

Compare it to the standard Gaussian distribution (that is computing the **Z-score**).

$$\frac{\bar{Y} - \mu}{\frac{7}{6}} = G(0, 1) = Z$$

Note: To get the Z score for the middle section contain 95%, we must look for the 2.5th and 97.5th percentile respectively!! Getting the Z score for 0.95, we have -1.96 and 1.96. Thus

$$P(-1.96 \leq Z \leq 1.96) = 0.95$$
$$P(-1.96 \leq \frac{\bar{Y} - \mu}{\frac{7}{6}} \leq 1.96) = 0.95$$

The left two terms of the inequality yield

$$\mu \leq \bar{Y} + 1.96\frac{7}{6}$$

and similar the right two terms yield

$$\mu \geq \bar{Y} - 1.96\frac{7}{6}$$

Thus our **coverage interval** is (95% chance of μ falling in this range)

$$P(\bar{Y} - 1.96\frac{7}{6} \leq \mu \leq \bar{Y} - 1.96\frac{7}{6}) = 0.95$$

So our estimate for this coverage interval

$$(\bar{y} - 1.96\frac{7}{6}, \bar{y} + 1.96\frac{7}{6})$$

Note that the **CONFIDENCE interval** would be for some $\bar{y}$. So if $\bar{y} = 80$, our confidence interval is

$$(80 - 1.96\frac{7}{6}, 80 + 1.96\frac{7}{6})$$

which is *based on the sample.*

More formally

Coverage Interval contains μ with 95% *probability*

Confidence Interval is the *estimate* of the coverage interval

Example. We want to find θ the approval rating of Trump, $n = 500$, $y = 220$ (220 approves Trump). Find the 95% CI (confidence interval) for θ.

Note that

$$Y \sim Bin(500, \theta)$$

where $\hat{\theta} = 220/500 = 0.44$ since $\hat{\theta} = y/n$ in a Binomial distribution. That is $\tilde{\theta} = Y/n$ (r.v. for θ).

By the *Central Limit Theorem (CLT)*,

$$\frac{\tilde{\theta} - \theta}{\sqrt{\frac{\tilde{\theta}(1-\tilde{\theta})}{n}}} = G(0, 1)$$

Thus we can use the Z scores

$$P(-1.96 \leq Z \leq 1.96) = 0.95$$

$$P(-1.96 \leq \frac{\tilde{\theta} - \theta}{\sqrt{\frac{\tilde{\theta}(1-\tilde{\theta})}{n}}} \leq 1.96) = 0.95$$

So our coverage interval is

$$(\tilde{\theta} \pm 1.96\sqrt{\frac{\tilde{\theta}(1-\tilde{\theta})}{n}}$$

and our confidence interval is

$$(0.44 \pm 1.96\sqrt{\frac{0.44 \times 0.56}{500}}$$

13 June 5, 2017

13.1 Interpretation of Confidence Interval

(i) It is the *estimate(s)* of the r.v.s L and U such that

$$P(L < \theta < U) = 0.95$$

(ii) If the experiment is repeated many times, and the CI constructed based on each sample, approximately 95% of them will fall within the range. ????

13.2 Steps to find Confidence Interval (CI)

Gaussian problem with known variance.

$Y_1, \ldots, Y_n$ are independent Gaussian with mean μ (unknown) and σ s.d. σ_0 (σ_0 known).

Our data is $\{y_1, \ldots, y_n\}$.

Objective: to construct a 90% CI for μ.

Step 1 Find the *estimate* of the unknown parameter. $\hat{\mu} = MLE = \bar{y} = \frac{1}{n}\sum y_i$. Note that $\bar{y}$ (estimate) is a known number.

Step 2 Identify the estimator and its distribution. $\bar{Y}$ (estimator) is the random variable from which $\bar{y}$ is an outcome. We know from STAT230 $\bar{Y} \sim G(\mu, \frac{\sigma_0}{\sqrt{n}})$ which is the sampling distribution of $\bar{Y}$.

Step 3 Construct the *Pivotal Quantity* (convert to something we know i.e. standard normal distribution) and find the pivotal distribution.

$$\frac{\bar{Y} - \mu}{\frac{\sigma_0}{\sqrt{n}}} = Z = G(0, 1)$$

where Z is the pivotal distribution and the LHS is the pivotal quantity.

Step 4 Find the end points (which depends on the level of confidence) of the pivotal distribution of step 3 (from percentile (of which p correspond to the proportion of data points within the percentiles) to Z score table). look for *0.90* (90% probability) to find that $x = 1.65$.

Step 5 Use step 4 to construct the *coverage interval*.

$$
\begin{aligned}
&P(-1.65 \leq Z \leq 1.65) = 0.9 \\
&P(-1.65 \leq \frac{\bar{Y} - \mu}{\frac{\sigma_0}{\sqrt{n}}} \leq 1.65) = 0.9 \\
&P(\bar{Y} - 1.65\frac{\sigma_0}{\sqrt{n}} \leq \mu \leq \bar{Y} + 1.65\frac{\sigma_0}{\sqrt{n}}) = 0.9.
\end{aligned}
$$

So our converage interval is $(\bar{Y} \pm 1.65\frac{\sigma_0}{\sqrt{n}})$.

Step 6 Estimate the coverage interval from your sample. Thus the cofidence interval is

$$[\bar{y} - 1.65\frac{\sigma_0}{\sqrt{n}}, \bar{y} + 1.65\frac{\sigma_0}{\sqrt{n}}]$$

In general for a Gaussian distribution with unknown variance, the CI for the mean is

$$[\bar{y} \pm z^* \frac{\sigma_0}{\sqrt{n}}]$$

where z^* depends on level of confidence.

13.3 Notes on Confidence Interval

- As n becomes *large*, the interval becomes *narrower* for the same level of confidence.
- As the level of confidence *goes up*, the interval will be *wider*.
- As σ_0 *goes up*, the interval will be *wider*.
- Can we choose the length of the CI? Suppose 95% CI to be length ± 5. That is

  $$\frac{z^*\sigma_0}{\sqrt{n}} = 5$$

  Thus we can choose n (number of samples) such that

  $$n = (\frac{z^*\sigma_0}{5})^2$$

  to make the CI have a length ± 5.

13.4 Fixing the Length of the Confidence Interval

Example. Binomial model where n is large. Note that $Y \sim Bin(n, \theta)$ where # of successes is y.

Based on data, construct a 95% CI for θ.

Step 1: the estimate $\hat{\theta} = y/n =$ sample proportion.

Step 2: the estimator $\tilde{\theta} = Y/n$. By CLT,

$$\tilde{\theta} \sim G(\theta, \sqrt{\frac{\tilde{\theta}(1-\tilde{\theta})}{n}}$$

Step 3: Pivotal quantity/distribution

$$\frac{\tilde{\theta} - \theta}{\sqrt{\frac{\tilde{\theta}(1-\tilde{\theta})}{n}}} = Z = G(0,1)$$

Step 4: Find points of pivotal distribution (pd)

$$P(-1.96 \leq Z \leq 1.96) = 0.95$$

Step 5: Coverage

$$P(-1.96 \leq \frac{\tilde{\theta} - \theta}{\sqrt{\frac{\tilde{\theta}(1-\tilde{\theta})}{n}}} \leq 1.96) = 0.95$$

So the coverage inteval is

$$\tilde{\theta} \pm 1.96\sqrt{\frac{\tilde{\theta}(1-\tilde{\theta})}{n}}$$

Step 6: Find the CI

$$\hat{\theta} \pm 1.96\sqrt{\frac{\tilde{\theta}(1-\tilde{\theta})}{n}}$$

Our margin of error is the $\pm$ term. Can we ensure the MOE ≤ 0.03, that is

$$1.96\sqrt{\frac{\tilde{\theta}(1-\tilde{\theta})}{n}} \leq 0.03$$

that is

$$n \geq (\frac{1.96}{0.03})^2 \times \hat{\theta}(1-\hat{\theta})$$

We choose n such that

$$n \geq (\frac{1.96}{0.03})^2 \times (\frac{1}{2})^2 \approx 1067$$

so one would only need to survey 1068 people to have a CI of $\pm 3\%$.

14 June 7, 2017

14.1 Pivotal Quantity

Definition. A pivotal quantity Q is a r.v. that depends on Y and θ such that $P(Q \geq a)$, $P(Q \leq b)$ for any a and b (can be calculated without knowing the value of θ).

14.2 Binomial Model Interval Estimation

Example. A sample of 1200 Americans are selected and 300 of them approve Trump. Find the $100q\%$ confidence interval for $\theta =$ population approval rating of Trump (note q is prespecified, e.g. $q = 0.9, 0.95, 0.99$).

Step 0: Set up the statistical model

$$Y \sim Bin(1200, \theta)$$

where θ is the probability of success.

Step 1: Find the estimate of θ (i.e. $\hat{\theta} = MLE$ where

$$\hat{\theta} = y/n = 300/1200 = 0.25$$

Step 2: Identify the estimator and the sampling distributions

$$\tilde{\theta} = Y/n$$

which is a r.v. for our estimate ($\bar{y}$) We know for the CLT, $\tilde{\theta} \sim Gaussian$

Step 3: Construct the pivotal quantity

$$\frac{\tilde{\theta} - \theta}{\sqrt{\frac{\tilde{\theta}(1-\tilde{\theta})}{n}}} = Z = G(0, 1)$$

where the LHS is Q, the pivotal quantity.

Step 4: Construct the interval for hte pivotal distribution For a 99% CI, From the Z-table, $Z^* = 2.58$

Step 5: Use Step 4 to construct the coverage interval

$$P(-2.58 < Z < 2.58) = 0.99$$

$$P(-2.58 < \frac{\tilde{\theta} - \theta}{\sqrt{\frac{\tilde{\theta}(1-\tilde{\theta})}{n}}} < 2.58) = 0.99$$

$$P(\tilde{\theta} - 2.58\sqrt{\frac{\tilde{\theta}(1-\tilde{\theta})}{n}} < \theta < \tilde{\theta} + 2.58\sqrt{\frac{\tilde{\theta}(1-\tilde{\theta})}{n}}) = 0.99$$

Step 6: Estimate the coverage interval to construct the CI The CI with $\hat{\theta} = 0.25$ and $n = 1200$ is

$$0.25 \pm 2.58\sqrt{\frac{0.25 \times 0.25}{1200}}$$

14.3 Sample Size

How to choose the "right" sample size?

Problem: Level of confidence (given 95%) and maximum length of interval is also prespecified

$$\hat{\theta} \pm l$$

where l is given. Can we choose n to make this happen?

Note that our CI for a binomial distribution is

$$\hat{\theta} \pm Z^* \sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}}$$

$$\rightarrow Z^* \sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}} \leq l$$

$$\rightarrow n \geq (\frac{Z^*}{l})^2 \hat{\theta}(1-\hat{\theta})$$

Choose n such that it is greater than the RHS for all values of $\hat{\theta}$. For example

$$max(\hat{\theta}(1-\hat{\theta})) = \frac{1}{2}\frac{1}{2} = \frac{1}{4}$$

Thus

$$n \geq (\frac{Z^*}{l})^2 \cdot \frac{1}{4}$$

We choose n to be the next biggest integer of the RHS. For example when we want a 95% CI and $l = 0.03$ or 3% margin of error

$$n \geq (\frac{1.96}{0.03})^2 \cdot \frac{1}{4} \approx 1068$$

For 90% confidence, we'd only need to survey 350 people.

14.4 Practical Surveys

Note articles involving surveys will always stat the MOE (e.g. ± 0.02 (l)) and the CI % (e.g. 19 times out of 20 = 95% CI).

Why not 99% CI? This would require surveying 3000 more people which takes more time which may allow $\hat{\theta}$ to change over time, skewing results.

14.5 Chi-Squared Distribution

Definition. Let W be a continuous r.v. W is said to be a **Chi-Squared distribution** with n **degrees of freedom** denoted by

$$W \sim X_n^2$$

if $W = Z_1^2 + Z_2^2 + \ldots + Z_n^2$ where $Z_i \sim G(0,1)$ and Z_i independent.

Geometrically for $W = Z_1^2 + Z_2^2$ a given data point (z_1, z_2) would be a plot on the Cartesian plane where Z_1 is the x-axis and Z_2 is the y-axis. The squared distance from the origin is called the **chi-squared distance**.

The possible values of $W \sim X_n^2$ is $W \in [0, \infty)$.
The special case is when $n = 1$.

Example. Suppose $W \sim X^2$. Find $P(W \leq 1.44)$. $n = 1 \to W = Z^2$. Thus

$$P(W \leq 1.44) = P(Z^2 \leq 1.44) = P(-1.2 \leq Z \leq 1.2)$$

14.6 Chi-Squared and CI

Example. Suppose $W \sim X^2$. Find the 95th percentile of W. Let the 95th percentile be x.

$$\begin{aligned}
&P(W \leq x) = 0.95 \\
&P(Z^2 \leq x) = 0.95 \\
&P(-\sqrt{x} \leq Z \leq \sqrt{x}) = 0.95
\end{aligned}$$

Note that $\sqrt{x} = 1.96$ (from Z-table), thus $x = 1.96^2$.

15 June 9, 2017

15.1 Chi-Squared Properties

1. W can take on any non-negative values $W \in [0, \infty)$
2. Degrees of freedom is a parameter n that specifies the number of Z_is. As n increases, the peak (median/mean) shifts to the right.
3. The expected value $E(W) = n$, the degrees of freedom.

   Proof. To show that $E(X) = n$, note that

   $$W = Z_1^2 + Z_2^2 + \ldots + Z_n^2$$

   so taking the expectation

   $$E(W) = E(Z_1^2) + E(Z_2^2) + \ldots + E(Z_n^2)$$

   Note that $Z_i \sim G(0, 1)$ thus $E(Z_i) = 0$ and $V(Z_i) = 1$. From the variance formula $V(Z_i) = E(Z_i^2) - (E(Z_i))^2 = 1$ thus $E(Z_i^2) = 1$. So $E(W) = n$. □

4. The variance is $V(W) = 2n$.

15.2 Addition of Chi Squared Distributions

Let $W_1 \sim X_{n_1}^2$ and $W_2 \sim X_{n_2}^2$, W_1, W_2 independent.

Let $W = W_1 + W_2$. What is the distribution of W? Then $W \sim X_{n_1+n_2}^2$.

In other words, the sum of two chi-squared distributions is a chi-squared distribution.

15.3 Probability Calculations

Note df = degrees of freedom.

Special Cases

Case 1 df = 1

If $n = 1$, then $W = Z^2$ where $Z \sim G(0, 1)$.

Example. Suppose $W \sim X_1^2$. Find c such that $P(W \leq c) = 0.85$. That is c is the 85th percentile.

$$\begin{aligned}
&P(W \leq c) = 0.85 \\
&\rightarrow P(Z^2 \leq c) = 0.85 \\
&\rightarrow P(-\sqrt{c} \leq Z \leq \sqrt{c}) = 0.85
\end{aligned}$$

Note the area inbetween $-\sqrt{c}$ and $\sqrt{c}$ is 0.85. Thus $\sqrt{c}$ is at the 0.925 (92.5th percentile) mark, which corresponds to a z-score of 1.44 thus $\sqrt{c} = 1.44$, so $c = 1.44^2$.

Case 2 df = 2

Result: If $W \sim X_2^2$, it is equivalent to saying $W = exp(\frac{1}{2})$ (their PDFs are the same).

Example. Suppose $W \sim X_2^2$. Find $P(W \leq 2.5)$.

$$P(W \leq 2.5) = \int_0^{2.5} \frac{1}{2} e^{-\frac{x}{2}} dx$$

Note for the exponential distribution, if $X \sim Exp(\lambda)$, then

$$F(\mu) = P(X \leq \mu) = 1 - e^{-\lambda\mu}$$

So for our example

$$P(W \leq 2.5) = 1 - e^{-\frac{2.5}{2}}$$

Case 3 df is "large" (> 30)

As n becomes large, the chi-squared approaches the Gaussian distribution with mean n and variance $2n$.

Example. Suppose $W \sim X_{72}^2$. Find $P(W \leq 96)$.

Note that $W \sim X_{72}^2 \rightarrow W \sim G(72, 12^2)$ (where $12 = \sqrt{2n}$). So we have

$$\begin{aligned}
P(W \leq 96) &= P(\frac{W - 72}{12} \leq \frac{96 - 72}{12}) \\
&= P(Z \leq 2)
\end{aligned}$$

which corresponds to a total probability of 0.95 (inside 2 sigma), thus $P(W \leq 96) = 0.95$.

Case 4 df lies between 2 and 30

We use the chi-squared table where the *rows = degrees of freedom* and the *columns = percentiles*. In row 15, column 0.025 we have 6.262. That is $P(W \leq 6.262) = 0.025$ for $W \sim X^2_{15}$.

Example. let $W \sim X^2_{17}$. Find a such that $P(W \geq a) = 0.05$.

We want to locate row 17 (df = 17) and column $1 - 0.05 = 0.95 \rightarrow a = 27.587$.

Example. let $W \sim X^2_{20}$. Find a, b such that $P(a \leq W \leq b) = 0.95$.

Note that a, b could be any interval as long as the CDF in the middle is 0.95 (a could be the 0.25th percentile and b the 97.5th, or a could be the 0th percentile and b could be the 95th).

The **convention** is to use the *equal-tailed* solution, so when a is the 0.25th percentile and b is the 0.975th percentile, or for row 20 (df = 20) we have $a = 9.591$ and $b = 34.170$.

16 June 12, 2017

16.1 T-Distribution

A random variable T is said to follow a **Student's T-distribution** with n degrees of freedom if T is a ratio of two independent r.v.s

$$T_n = \frac{Z}{W}$$

where Z, W are independent and

$$Z \sim G(0,1)$$
$$W = \sqrt{\frac{X^2(n)}{n}}$$

where $X^2(n)$ is the n degree chi-squared distribution.

16.2 Properties of T-Distribution

1. $T_n \in (-\infty, \infty)$ for all n
2. T_n is symmetric around zero $\forall n$, that is mean = median = 0
3. T is "similar" in shape to the Z-distribution but T has *fatter tails* (more extreme observations compared to the Z-distribution). THis means $K > 3$ for any n
4. As $n \rightarrow \infty$, $T_n \rightarrow Z$ (The T-distribution approaches the Z-distribution as $n \rightarrow \infty$)

16.3 Student T Table

Similar to a chi-squared table.

Example. Suppose T is a r.v. which follows a T-distribution with $n = 23 = df$. Find a number c such that

$$P(|T| \leq c) = 0.95$$

$$\begin{aligned}&P(|T_{23}| \leq c) = 0.95\\ &\to P(-c \leq T_{23} \leq c) = 0.95\end{aligned}$$

We look up row 23 and column 0.975 to find c (since T-distribution is symmetric, we need not look up 0.025). So $c = 2.0687$.

Theorem. Let $Y_1, \ldots, Y_n$ be independent Gaussian random viarables with mean μ and variance σ^2 (where μ, σ^2 are unknown). That is our sample is $\{y_1, \ldots, y_n\}$. How do we figure out μ from just our sample (and its sample mean $\bar{y}$ and sample variance s^2.

Note $\bar{Y} = \frac{1}{n} \sum Y_i$ is the estimator corresponding to the sample mean.

Note $S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (Y_i - \bar{Y})^2$ is the estimator corresponding to the sample variance.

Then our theorem is:

(a) To estimate μ from $\hat{\bar{y}}$ and s, use

$$\frac{\bar{Y} - \mu}{\frac{s}{\sqrt{n}}} \sim T_{n-1}$$

(b) To estimate σ^2 from $\bar{y}$ and s, use

$$\frac{(n-1)S^2}{\sigma^2} \sim X^2_{n-1}$$

where $n - 1$ are the degrees of freedom, s the sample deviation, and S^2 the estimator for sample variance.

Note $\sigma^2 \neq s^2 \neq S^2$, where S^2 is the random variable from which s^2 is drawn. Note that σ^2 is an unknown constant.

Note that (a) is relevant when we have $Y = Y_1 + \ldots + Y_n$ where $Y_i \sim G(\mu, \sigma^2)$ since $Y \sim G(\mu, \frac{\sigma}{\sqrt{n}}^2)$ or

$$\frac{Y - \mu}{\frac{\sigma}{\sqrt{n}}} \sim Z$$

for a known σ (but unknown μ). Remember this was our pivot quantity for the pivotal distribution of a Gaussian distribution.

Proof. Assume (b) is true. Is (a) true? We can rewrite (a) in the form

$$\begin{aligned}\frac{\bar{Y}-\mu}{\frac{s}{\sqrt{n}}} &= \frac{\frac{\bar{Y}-\mu}{\frac{\sigma}{\sqrt{n}}}}{\sqrt{\frac{(n-1)S^2}{\sigma^2}\cdot\frac{1}{n-1}}} \\ &= \frac{Z}{\sqrt{X_{n-1}^2\cdot\frac{1}{n-1}}} \\ &\frac{Z}{W}\end{aligned}$$

as desired. □

16.4 Expectation of S^2

For S^2, note that $E(S^2) = \sigma^2$ (**the average of our sample variance is the population variance**). This follows from (note $E(X_k^2) = k$).

$$\begin{aligned}E(\frac{(n-1)S^2}{\sigma^2}) &= n-1 \\ \frac{(n-1)}{\sigma^2}E(S^2) &= n-1 \\ E(S^2) &= \sigma^2\end{aligned}$$

this is why we divide by $n-1$ and not n. This is to say S^2 is an *unbiased estimator* of σ^2.

16.5 Unknown Mean μ from sample variance s (T-distribution)

Example. Suppose the income of Waterloo residents are Gaussian with mean μ and s.d. σ. A sample of size 20 is drawn where $\bar{y} = 50000$ and $s = 5000$. Based on the data, what is the 95% CI for μ?

Step 1: Find the estimate of mu

$$\hat{\mu} = \bar{y} = \frac{1}{n}\sum y_i$$

Step 2: Identify the estimator $\bar{Y}$

Step 3: Construct the pivot (from part (a) of our theorem), that is

$$\frac{\bar{Y}-\mu}{\frac{s}{\sqrt{n}}} \sim T_{19}$$

Step 4: Find the endpoints of the pivot We want $P(-c < T_{19} < c) = 0.95$. We want row 19 (**note df = n-1, NOT n**) and column 0.975, that is $c = 2.093$

Step 5: Find the coverage interval

$$P(-2.093 \leq T_{19} \leq 2.093) = 0.95$$
$$P(-2.093 \leq \frac{\bar{Y} - \mu}{\frac{s}{\sqrt{n}}} \leq 2.093) = 0.95$$
$$P(\bar{Y} - 2.093 \cdot \frac{s}{\sqrt{n}} \leq \mu \leq \bar{Y} + 2.093 \cdot \frac{s}{\sqrt{n}}) = 0.95$$

Step 6: The CI (confidence) would be

$$\bar{y} \pm 2.093 \frac{s}{\sqrt{n}} = 50000 \pm 2.093 \cdot \frac{5000}{\sqrt{20}}$$

Key note: For small sample sets, the T-distribution is useful. For really large sample sets, $n \to \infty$ which approaches a Gaussian distribution anyways.

17 June 14, 2017

17.1 Unknown Variance σ from Sample Variance s (Chi-Squared)

What if we wanted to calculate the unknown σ from the example before? Use part (b) of our theorem

Example. Suppose the income of Waterloo residents are Gaussian with mean μ and s.d. σ. A sample of size 20 is drawn where $\bar{y} = 50000$ and $s = 5000$. Based on the data, what is the 95% CI for σ?

Step 1: The pivotal quantity is

$$\frac{14S^2}{\sigma^2} \sim X^2(14)$$

where the RHS is the **pivotal distribution**.

Step 2: Construct coverage interval. Note a and b are found from Chi-Squared table (where the convention is to have equal tail intervals (so at a where $p = 0.025$ and b where $p = 0.975$, and $df = 14$).

$$P(a \leq X^(14) \leq b) = 0.95$$
$$P(a \leq \frac{14S^2}{\sigma^2} \leq b) = 0.95$$

So we get

$$\sigma^2 \geq \frac{14S^2}{b}$$
$$\sigma^2 \leq \frac{14S^2}{a}$$

that is

$$P(\frac{14S^2}{b} \leq \sigma^2 \leq \frac{14S^2}{a}) = 0.95$$

Therefore our CI is

$$(\frac{14s^2}{b}, \frac{14s^2}{a})$$

where s^2 is our best estimate of S^2 or simply the sample variance. a and b are from the Chi-Squared table.

More generally, the CI for the variance σ^2 is

$$(\frac{(n-1)s^2}{b}, \frac{(n-1)s^2}{a})$$

17.2 CI for Poisson

Large n

Given $Y \sim Pois(\theta)$, note that $\mu = \theta$ and $\sigma^2 = \theta$. For a **large enough** $n > 30$ sample size, by the CLT

$$\frac{\bar{Y} - \theta}{\sqrt{\frac{\bar{y}}{n}}} \sim G(0, 1)$$

this holds because $V(\bar{Y}) = \frac{\theta}{n}$ so $SD(\bar{y}) = \sqrt{Var(\bar{Y}} = \sqrt{\frac{\theta}{n}}$, and this is the pivotal quantity for a Gaussian distribution with denominator $\frac{\sigma}{n}$.

Solving for the CI for θ, we get the general form for Poisson distributions (with large n)

$$\bar{y} \pm z^* \sqrt{\frac{\bar{y}}{n}}$$

17.3 CI for Exponential

Given $\bar{Y} = Y_1, \ldots, Y_n \sim Exp(\theta)$, note $E(Y_i) = \theta$ and $V(Y_i) = \theta^2$. Therefore $E(\bar{Y}) = \theta$ and $V(\bar{Y}) = \frac{\sigma^2}{n}$ so $SD(\bar{Y}) = \frac{\sigma}{\sqrt{n}}$.

Thus by CLT

$$\bar{Y} \sim G(\theta, \frac{\theta}{\sqrt{n}})$$

where we have the pivotal relation

$$\frac{\bar{Y} - \theta}{\frac{\theta}{\sqrt{n}}} = G(0, 1)$$

Small n

If $Y \sim Exp(\theta)$, then

$$\frac{2Y}{\theta} \sim Exp(2)$$

This follows because if you divide every point in the exponential distribution by the mean θ (normalized), then multiply by 2, then it should be an exponential distribution with mean 2. Furthermore, note that

$$Exp(2) \sim X^2(2)$$

Adding up n of these we get

$$\frac{2}{\theta}\sum^{n} Y_i \sim X^2(2n)$$

18 June 16, 2017

18.1 Exponential Example

The lifetime of a light bulb has an Exponential distribution with mean θ.

A sample of observations are drawn $\{y_1, \ldots, y_n\}$. Note that $\bar{y} = 10000$. Find the 95% CI for θ.

Step 1: Find the estimate of θ. Note that $\hat{\theta} = \bar{y}$.

Step 2: Identify the estimator: $\bar{Y}$ (Sums of exponential distribution)

Step 3: Find the pivotal quantity and identify the pivotal distribution

$$W = \frac{2}{\theta}\sum Y_i = X^2(30)$$

Step 4: Find the end points of the pivotal distribution, that is find a and b such that

$$P(a < X^2(30) < b) = 0.95$$

Looking up row 30 where a is column 0.025 and b is column 0.975 we get $a = 16.791$ and $b = 46.979$.

Step 5: Find the coverage interval

$$P(16.791 \leq X^2(30) \leq 46.979) = 0.95$$
$$P(16.791 \leq \frac{2}{\theta}\sum Y_i \leq 46.979) = 0.95$$

So $\theta \geq \frac{2\sum Y_i}{46.979}$ and $\theta \leq 2\frac{2\sum Y_i}{16.791}$.

Thus we have the coverage interval

$$[\theta\frac{2n\bar{y}}{46.979}, \theta\frac{2n\bar{y}}{16.791}]$$

In general we have

$$[\frac{2n\bar{y}}{b}, \frac{2n\bar{y}}{a}]$$

where a and b are computed from the $X^2(2n)$ table.

18.2 LI vs CI

Note that $100p\%$ LI for $\theta = \{\theta : R(\theta) \geq p\}$.

18.3 Likelihood Ratio Test Statistic

Theorem. If θ is the true value of the unknown parameter and $\hat{\theta} = MLE$ then, for a large n

$$\Lambda(\theta) = -2ln\frac{L(\theta)}{L(\tilde{\theta})} \sim X^2(1)$$

where $\tilde{\theta}$ is the estimator corresponding to the MLE.

Consider $\Lambda(\theta) = -2ln\frac{L(\theta)}{L(\hat{\theta})}$: these are all outcomes of the $X^2(1)$ distribution if n is large.

We call Λ the **Likelihood Ratio Test Statistic**.

18.4 Confidence to Likelihood

Example. Suppose n is large, and we have a 95% CI. What likelihood interval does this correspond to?

$$\begin{aligned}
&P(-1.96 < Z < 1.96) = 0.95 \\
\iff\ &P(Z^2 \leq 1.96^2) = 0.95 \\
\iff\ &P(\Lambda(\theta) \leq 1.96^2) = 0.95 \\
\iff\ &P(-2ln\frac{L(\theta)}{L(\tilde{\theta})} \leq 1.96^2) = 0.95 \\
\iff\ &P(\frac{L(\theta)}{L(\tilde{\theta})} \geq e^{\frac{-1.96^2}{2}}) = 0.95
\end{aligned}$$

where this resembles $\{\theta : R(\theta) \geq p\}$, thus $p = e^{\frac{-1.96^2}{2}} = 0.146$.

So a 95% CI is approximately a 14.6% LI. Similar for a 90% CI, it is approximately a $e^{\frac{-1.64^2}{2}}$ LI.

In general

$$100p\%CI = e^{\frac{-z^{*2}}{2}}$$

where z^* can be computed from the Z-table.

18.5 Likelihood to Confidence

Example. Suppose we have a 10% LI. What is the corresponding CI?

$$
\begin{aligned}
&R(\theta) \geq 0.1 \\
&\frac{L(\theta)}{L(\hat{\theta})} \geq 0.1 \\
&-2ln\frac{L(\theta)}{L(\hat{\theta})} \leq -2ln(0.1)
\end{aligned}
$$

So the coverage interval is

$$
\begin{aligned}
&P(-2ln\frac{L(\theta)}{L(\tilde{\theta})} \leq -2ln(0.1)) \\
&= P(Z^2 \leq -2ln(0.1)) \\
&= P(-\sqrt{-2ln(0.1)} \leq Z \leq \sqrt{-2ln(0.1)})
\end{aligned}
$$

Note $ln(0.1) = -2.14$, so we look at z-score $= 2.14$ which corresponds to a 96.8% CI.

Note a 50% LI = 76% CI. A 1% LI = 99% CI.

So a wider LI corresponds to a narrow CI. A narrower LI correspond to a wider CI.

18.6 Prediction Interval

Given $Y_1, Y_2, \ldots, Y_n$ are independent Gaussian r.v.s with mean μ and s.d. σ.

Note we use time series analysis if i stands for time.

Data set is $\{y_1, \ldots, y_n\}$.

Objective: To construct a 95% prediction interval for Y_{n+1}.

Some examples include:

- Birth rates in Canada where Y_i is the # of children born in Canada in month i. Based on data, we want to predict Y_{n+1}.
- Job Market: forecasting future qualities of candidates can help select the right candidate.
- Y_i = stock price of a company in time i. We have data for the past in periods. We want to predict Y_{n+1}.

How to find prediction interval:

Note $Y_1, \ldots, Y_n, \ldots$ are all Gaussian (μ, σ). So $Y_i \sim G(\mu, \sigma)$. Note that

$$\bar{Y} \sim G(\mu, \frac{\sigma}{\sqrt{n}})$$

where $\bar{Y}$ is the sample mean r.v. Furthermore, note

$$Y_{n+1} \sim G(\mu, \sigma)$$

Are $\bar{Y}$ and Y_{n+1} independent? Yes they are.

Question: What distribution does $Y_{n+1} - \bar{Y}$ follow? Note that if X and Y are independent Gaussian where $X \sim G(\mu_1, \sigma_1)$ and $Y \sim G(\mu_2, \sigma_2)$ then

$$aX + bY \sim G(a\mu_1 + b\mu_2, \sqrt{a^2\sigma_1^2 + b^2\sigma_2^2})$$

In the example, $a = 1, b = -1$ thus we get

$$Y_{n+1} - \bar{Y} \sim G(0, \sqrt{\sigma^2 + \frac{\sigma^2}{n}})$$
$$Y_{n+1} - \bar{Y} \sim G(0, \sigma\sqrt{1 + \frac{1}{n}})$$

Extracting the pivotal quantity with pivotal distribution as the Z distribution

$$\frac{Y_{n+1} - \bar{Y}}{\sigma\sqrt{1 + \frac{1}{n}}} \sim Z(0, 1)$$

Note however we do not know what σ is!

We can use the sample s.d. s instead with pivotal distribution as the T distribution

$$\frac{Y_{n+1} - \bar{Y}}{s\sqrt{1 + \frac{1}{n}}} \sim T_{n-1}$$

Example. Suppose $n = 20$. Go to the T-Table and find t^* (as $n-1$) (for 95%)

$$P(-t^* \leq T \leq t^*) = 0.95$$

Note $t^* = 2.093$, thus we get

$$P(-2.093 \leq \frac{Y_{n+1} - \bar{Y}}{s\sqrt{1 + \frac{1}{n}}} \leq 2.093) = 0.95$$
$$P(\bar{Y} - 2.093s\sqrt{1 + \frac{1}{n}} \leq Y_{n+1} \leq \bar{Y} + 2.093s\sqrt{1 + \frac{1}{n}}) = 0.95$$

Thus the prediction interval for Gaussian r.v.s is

$$\bar{y} \pm t^* s\sqrt{1 + \frac{1}{n}}$$

Note we had assumed Y_is are all independent (which is almost never true for time series data).

18.7 Testing of Hypotheses

A **hypothesis** is a statement made about some parameter of the population. Ex: $\theta = \theta_0$

This statement can only be checked if we had the *entire population*.

We can take a sample and based on the sample we decide whether or not the hypothesis holds.

18.8 Null and Alternate Hypothesis

There are two competing hypotheses

Null Hypothesis (H_0) This is the conventional wisdom (the current belief).

Alternate Hypothesis(H_1, H_A) This is the challenger to the current belief.

18.9 Analogy to Legal System

Testing is very similar to the legal system. For example, the null hypothesis H_0 for a suspect in trial is that they are innocent. H_1 is that they are guilty. In a trial, we assume the suspect is innocent until proven guilty. Based on evidence, we *convict* when we reject H_0 when there is enough evidence, and we acquit when we do not reject H_0 when there is enough evidence, and we acquit when we do not reject H_0

That is the *lower the p-value, the stronger the evidence* against H_0. Typically, we choose the p-value $0.05, 0.01$, etc.

19 June 21, 2017

19.1 Examples of Null and Alternate Hypotheses

Example. Jeopardy

Let θ be the probability that a Canadian wins Jeopardy. Let our claim (alternate hypothesis) H_1 be that $\theta > \frac{1}{3}$. Note the status quo or null hypothesis is H_0 where $\theta = \frac{1}{3}$.

Example. Discrimination

Is there discrimination against women in salary terms? Let μ_1 be the average salary of men, μ_2 for women. Note that the null hypothesis H_0 would be that $\mu_1 = \mu_2$ (status quo that there is no discrimination). H_1 is $\mu_1 > \mu_2$.

19.2 p-value

Definition. The **p-value** is the probability of observing your evidence (in the form of a test statistic d) (or worse/more extreme) *given that* H_0 *is true.*

This *does not mean* that the probability of H_0 is true is 0.3 for $p = 0.3$. It means among all H_0 cases, 0.3 of the population exhibits the evidence. That is, how unusual the evidence is amongst the population of H_0.

The lower value the p-value, the stronger the evidence against H_0.

19.3 Convention for p-value

$p > 0.1$ No evidence against H_0

$0.05 < p \leq 0.1$ Weak evidence

$0.01 < p \leq 0.05$ Strong evidence

$p \leq 0.01$ Very strong evidence

Typical the cut-off is 5% or when $p = 0.05$. That is when $p \leq 0.05$, then we reject H_0, otherwise we do not reject H_0.

19.4 Type of Errors in Hypothesis Testing

Type I error When you **reject** H_0 when it is in fact true.

Type II error When you **do not reject** H_0 when it is in fact false.

Note Type I error is the more errorneous error (e.g. convicting an innocent defendant is worse than not convicting a criminal).

We want to decide on tests with a low type I error probability (< 0.05).

20 June 23, 2017

20.1 Hypothesis Testing Example

Example. Test whether a coin is fair. A coin is tossed 20 times. $Y = \#ofheads$. Note H_0 is when $\theta = \frac{1}{2}$ and H_1 is when $\theta \neq \frac{1}{2}$.

This is a **two-tailed/sided tests:** both "high" and "low" values are bad news for H_0.

The Roll up the Rim is an example of a **one-sided test** where $H_0 = \frac{1}{6}$ and $H_1 < \frac{1}{6}$.

We will focus on two-sided tests.

We can construct a *discrepancy measure (D)*.

20.2 Discrepancy Measure

Definition. A discrepancy of a r.v. which measures the *level of disagreement* between the data and H_0.

Typically D satisfies the following properties

(i) $D \geq 0$

(ii) $D = 0$ implies best evidence for H_0

(iii) Larger the value of D, stronger evidence against H_0

(iv) $P(D \geq d)$ (p-value) can be calculated if H_0 is true

Example. Going back to the fair coin example, $H_0 : \theta = \frac{1}{2}$ and $H_1 : \theta \neq \frac{1}{2}$ where $\theta = P(H)$.

If $n = 200$, then is $D|Y - 100|$, where Y is the number of heads, a good D?

If the p-value ($P(D \geq d)$) is small means one of two things:

(i) H_0 is true, and you observed a *really rare event*

(ii) H_0 is not true

Example. $Y_1, \ldots, Y_{25}$ are iid Gaussian r.v.s with mean μ and sd σ. Let our sample $\{y_1, \ldots, y_{25}\}$ with $\bar{y} = 8$.

Let $H_0 : \mu = 6$ and $H_1 : \mu \neq 6$. What should we conclude?

Note the pivotal quantity is

$$\frac{\bar{Y} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

If we assume H_0 is true (we do), then we can use

$$D = |\frac{\bar{Y} - 6}{\frac{\sigma}{\sqrt{n}}}|$$

which maps to the standard normal distribution which is 0 when $\mu = 6$ (we assumed this for H_0.

20.3 Calculate d and p-value

In the above example, note $\bar{y} = 8$. Suppose $\sigma = 5$. Plugging this into D, we see that

$$d = |\frac{8 - 6}{\frac{5}{\sqrt{25}}}| = 2$$

So we have $P(D \geq d) = P(|Z| \geq 2)$. Looking this up in the Z table, we see that the p-value < 0.05.

Therefore there is strong evidence against H_0.

20.4 When σ is Unknown

Example. $Y_1, \ldots, Y_{25} \sim G(\mu, \sigma)$. Note that our hypotheses are $H_0 : \mu = 6$ and $H_1 : \mu \neq 6$.

Finally let $\bar{y} = 8$ and $s^2 = 25$ and σ unknown.

How do we calculate the p-value? Use the T *distribution*.

$$D = |\frac{\bar{Y} - 6}{\frac{s}{\sqrt{n}}}|$$

Which is T_{24}. Note that

$$d = |\frac{8 - 6}{\frac{5}{5}} = 2$$

So we have $P(|T_{35}| \geq 2)$. So p-value is in between 5 and 10%, which is weak evidence against H_0.

20.5 Binomial Example

Coin question. $H_0 : \theta = \frac{1}{2}$ and $H_1 : \theta \neq \frac{1}{2}$.

Let $n = 200$. Our pivotal quantity (from CLT) is

$$\frac{\tilde{\theta} - \theta}{\sqrt{\frac{\theta(1-\theta)}{n}}}$$

so

$$D = |\frac{\tilde{\theta} - 0.5}{\sqrt{\frac{0.5^2}{n}}}|$$

21 June 26, 2017

21.1 Summary of Hypothesis Testing

Step 0 Set up the model

Step 1 Set up the null and alternate hypotheses

$$H_0 : \theta = \theta_0$$
$$H_1 : \theta \neq \theta_0$$

where θ_0 is a given constant, θ unknown

Step 2 Construct the discrepancy measure D (Test-Statistic r.v.) and calculate d (value of test-statistic in your sample, outcome of r.v. D). Some desirable properties of D:

1. $D \geq 0$
2. $D = 0 \rightarrow$ best evidence for null hypothesis
3. $D >> 0 \rightarrow$ strong evidence against H_0
4. $P(D \geq d)$ can be calculated if we assume that H_0 is true

Step 3 Calculate the p-value using $d =$ outcome from your sample

$$\text{p-value } = P(D \geq d; H_0 \text{ is true})$$

Note this will most likely be two-tailed since $D = |Z|$.

Step 4 Based on Step 3, we draw appropriate conclusions

The conclusion of a test can be written in one of two ways:

1. Find the p-value and apply the chart provided
2. We decide on a cut-off p-value (e.g. $p = 0.05$). If $p < 0.05$ then we reject H_0 at 5% level of confidence. Otherwise ($p \geq 0.05$) we do not reject H_0 at 5% level.

   In the social sciences, $p = 0.05$ typically. Physical sciences $p = 0.01$ (usually).

21.2 Sigma Test

Note that a p-value of 0.05 corresponds to a 2-sigma test. Similarly, a $1-p(n\sigma)$ p-value corresponds to an n-sigma test (1-sigma test has a p-value of $1-p(1) = 1-0.68 = 0.32$).

21.3 Another Hypothesis Testing Example

Example. $Y_1, \ldots, Y-n \sim G(\mu, \sigma)$ independent, where μ, σ unknown.

Note that $n = 16, \bar{y} = 7 \cdot 8, s = 4$.

We want to test whether $H_0 : \mu = 10$ (μ_0) or $H_1 : \mu \neq 10$.

Solution: Construct the test-statistic

$$D = |\frac{\bar{Y} - \mu_0}{\frac{S}{\sqrt{n}}}| = |\frac{\bar{Y} - 10}{\frac{S}{\sqrt{n}}}|$$

and we calculate d

$$d = |\frac{\bar{y} - 10}{\frac{s}{\sqrt{n}}}| = 2.2$$

Next we calculate the p-value where

$$\begin{aligned} \text{p-value } &= P(D \geq d) \\ &= P(|T_{15}| \geq 2.2) \end{aligned}$$

Note we have a two-tailed problem (absolute sign; need to find area of PDF in between ± 2.2).

From the T-table, we see that there is a large gap for 2.2 (between $p = 0.975$ and $p = 0.99$). Thus we know $0.01 < P(T_{15} \geq 2.2) < 0.025$ so (with the two tails) p-value must be between 0.02 and 0.05. There is thus strong evidence against H_0.

21.4 Relationship between CI and p-value

Theorem. If θ_0 belongs to the $100q\%$ CI, where $q \in (0, 1)$, the p-value of the test

$$\begin{aligned} H_0 &: \theta = \theta_0 \\ H_1 &: \theta \neq \theta_0 \end{aligned}$$

must be bigger than $1 - q$ (assuming we use the same pivot).

Proof. Proof for Gaussian:

Gaussian μ with known σ.

$$\begin{aligned} H_0 &: \mu = \mu_0 \\ H_1 &: \mu \neq \mu_0 \end{aligned}$$

If p-value > 0.05, then $\mu_0 \in 95\%$ CI.

$$P(D \geq d) > 0.05$$
$$P(|\frac{\bar{\mu} - \mu_0}{\frac{\sigma}{\sqrt{n}}}| \geq d) > 0.05$$
$$P(|Z| \geq d) > 0.05$$

So $d < 1.96$ (since we want the middle portion to be smaller than 0.95). Therefore μ_0 must belong to the 95% CI. $\square$

21.5 One vs Two Sided Tests

For Roll up the Win hypothesis testing, note that instead of a two sided test

$$H_0 : \theta = \frac{1}{i}$$
$$H_1 : \theta \neq \frac{1}{i}$$

where $\theta = P(prize)$, we have a one-sided inequality

$$H_0 : \theta = \frac{1}{i}$$
$$H_1 : \theta < \frac{1}{i}$$

Example. Let $n = 300$, $Y = \#ofprizes$. If $Y > 50$, then $D = 0$ (right-tail is irrelevant): best news for null hypothesis. Otherwise is $Y < 50$, then

$$D = |\frac{\tilde{\theta} - \theta_0}{\sqrt{\frac{\theta_0(1-\theta_0)}{n}}}|$$

21.6 Poisson Example

Example. Assume n is large, $Y_1, \ldots, Y_n \sim Pois(\mu)$

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

By CLT

$$D = |\frac{\bar{Y} - \mu_0}{\sqrt{\frac{\mu_0}{n}}}| = Z$$

with the test-statistic sample being

$$d = |\frac{\bar{y} - \mu_0}{\sqrt{\frac{\mu_0}{n}}}|$$

The p-value will thus be

$$= P(D \geq d)$$
$$= P(|Z| \geq d)$$

22 June 28, 2017

22.1 Poisson Testing

Example. Sample $\{y_1, \ldots, y_50\}$ and $H_0 : \mu = 5, H_1 : \mu \neq 5$. Also $\bar{y} = 6$. What can we conclude?

Construct the test statistic D and calculate outcome from experiment d:

$$D = |\frac{\bar{Y} - \mu}{\sqrt{\frac{\mu}{n}}} = |\frac{\bar{Y} - 5}{\sqrt{\frac{5}{n}}}$$

Thus our d is

$$d = |\frac{6 - 5}{\sqrt{\frac{5}{50}}} = \sqrt{10} \approx 3.1$$

Finally the p-value is

$$\begin{aligned} P(D \geq d) &= p(|Z| \geq 3.1) \\ &< 0.01 \end{aligned}$$

very strong evidence against H_0.

22.2 Measurement Bias Testing

Note that **bias is NOT the same as accuracy**.

Example. Take an object of known weight and measure it using your scale n times where Y_i is your ith reading on your scale and δ is the bias of your scale. Thus we have

$$Y_i = 10 + \delta + R_i$$

where R_i is the error in measurement, $R_i \sim G(0, \sigma)$.

We would like to test $H_0 : \sigma = 0, H_1 : \sigma \neq 0$.

Note $10+\delta$ is just a constant, so $Y_i = \mu+R_i$ is in fact a Gaussian distribution $G(\mu, \sigma)$ where $\mu = 10 + \delta$. Note μ is called the **systematic part** and R_i is the **random part** of the model.

Thus our hypotheses are now $H_0 : \mu = 10, H_1 : \mu \neq 10$.

Suppose $n = 36, \bar{y} = 13, s = 12$. We have the test statistic D

$$D = |\frac{\bar{Y} - \mu_0}{\frac{S}{\sqrt{n}}}| = |\frac{\bar{Y} - 10}{\frac{s}{\sqrt{n}}}|$$

Thus d is

$$d = |\frac{13 - 10}{\frac{12}{\sqrt{36}}}| = 1.5$$

So our p-value is

$$P(D \geq d) = P(|T_{35}| \geq 1.5) > 0.05$$

We do not have enough evidence to reject H_0.

Note there were two factors that affected our results in the example:

1. n sample size (If we change $n = 36$ to $n = 1000$, our d goes up and we reject H_0).

2. s variability (If we change $s = 12$ to $s = 1.2$, note our d goes up and thus we will reject H_0).

22.3 Testing for Variance

Let $Y_1, \ldots, Y_n \sim G(\mu, \sigma)$ independent. We want to test

$$H_0 : \sigma^2 = \sigma_0^2$$
$$H_1 : \sigma^2 \neq \sigma_0^2$$

Note we have the distribution with σ (and thus test statistic D)

$$D = \frac{(n-1)S^2}{\sigma^2} \sim X^2_{n-1}$$

Issues with Chi-Squared as Test Statistic

Note $D \geq 0$, but it does not satisfy the desired property where $D = 0$ is best evidence for H_0. The baseline is no longer 0 but $n - 1$ (expected value of X^2_{n-1}). $D >> n - 1$ and $D << n - 1$ are evidences against H_0.

X^2 is not symmetric.

The convention is:

(a) If d is right of the median of X^2, then the p-value $= 2P(D \geq d)$

(b) If d is left of median, then the p-value $= 2P(D \leq d)$

This bounds p-value to be ≤ 1.

Example. Suppose $Y_1, \ldots, Y_{51} \sim G(\mu, \sigma)$. Let our sample be $n = 51, \bar{y} = 10, s^2 = 2.055$. Furthermore we'd like to test $H_0 : \sigma^2 = 1, H_1 : \sigma^2 \neq 1$.

Our test statistic D is

$$D = \frac{(n-1)S^2}{\sigma_0^2} = \frac{(n-1)S^2}{1}$$

So our d from measurement is

$$d = (50)(2.055) = 102.75$$

Note X^2_{50} has $\mu = 50$ and $\sigma^2 = 100$ (also X^2 is roughly normal). Thus we know d lies far right of the median. So the p-value is $2P(D \geq 102.75)$.

22.4 More Statistics about Chi-Squared

For a given $X^2(n)$ distribution, we have

mean $= n$

mode $= n - 2$

median $\approx n - 0.7$

The median tells us whether d lies to the right/left of the median.

23 June 30, 2017

23.1 Degrees of Freedom

Given $Y_i = \mu + R_i$ (from e.g. systemic bias model), the *degrees of freedom* is the number of samples subtract the number of unknowns in the systematic part of the model. Since there is one unknown in Y_i, then $df = n - 1$.

23.2 Testing with Likelihood Function

Given a sample distribution

$$Y_i \sim f(y_i; \theta)$$

where Y_i independent, θ is an unknown parameter, we would like to test

$$H_0 : \theta = \theta_0$$
$$H_1 : \theta \neq \theta_0$$

In some cases, constructing D might be difficult if we do not know the properties of the distribution. (e.g. unknown distribution).

We can use the *LRTS (likelihood ratio test statistic)*

$$D = \Lambda(\theta_0) = -2ln\frac{L(\theta_0)}{L(\tilde{\theta})}$$

where $\tilde{\theta}$ is the MLE. Thus our measurement for a given observation is

$$d = \lambda(\theta_0) = -2ln\frac{L(\theta_0)}{L(\hat{\theta})}$$

Note D is equivalent to X_1^2, which indeeds satisfies the properties:

(i) $\Lambda \geq 0$

(ii) $\Lambda = 0$ best case for H_0

(iii) $\Lambda >> 0$ evidence against

(iv) Distribution $H_0 = X^2(1)$

Example. Is a coin fair? Toss coin 200 times where Y = number of heads.

Our sample yields $y = 110$. We want to test $H_0 : \theta = 0.5, H_1 : \theta \neq 0.5$ where $\theta = P(H)$. What should we conclude?

Step 1: Set up hypotheses, done!

Step 2: Set up D and calculate d

$$\Lambda = -2ln\frac{L(\theta_0)}{L(\tilde{\theta})}$$

so

$$\lambda = -2ln\frac{L(\theta_0)}{L(\hat{\theta})}$$

Note that the likelihood function for our binomial distribution is

$$L(\theta) = \binom{200}{110}\theta^{110}(1-\theta)^{90}$$

Note that $\hat{\theta} = 110/200 = 0.55$ and $\theta_0 = 0.5$. If we plug in our numbers we get $\lambda = 2.003$.

Step 3: Calculate the p-value

$$\begin{aligned} \text{p-value} &= P(D \geq d) \\ &= P(\Lambda \geq \lambda) \\ &= P(\Lambda \geq 2.003) \\ &= P(Z^2 \geq 2.003) \\ &\approx 0.16 \end{aligned}$$

where this was calculated by taking $P(Z \leq -\sqrt{2.003}) + P(Z \geq \sqrt{2.003})$.

Step 4: Draw appropriate conclusion. Since $p = 0.16$, we have no evidence against H_0.

Example. $Y_1, \ldots, Y_{50} \sim Exp(\theta)$ and

$$\begin{aligned} H_0 &: \theta = 2000 \\ H_1 &: \theta \neq 2000 \end{aligned}$$

A sample of 50 observations yields $\bar{y} = 1867.8$ hours.

The traditional method is to use the $X^2(2n)$ pivot.

Instead we can use

$$\Lambda = -2ln\frac{L(\theta_0)}{L(\tilde{\theta})}$$

so

$$\lambda = -2ln\frac{L(\theta_0)}{L(\hat{\theta})}$$

Note $L(\theta) = \frac{1}{\theta} e^{-\frac{y}{\theta}}$. Note $\hat{\theta} = 1867.8$ and $\theta_0 = 2000$, so we get $z = 0.1979$. Thus

$$
\begin{aligned}
p &= P(\Lambda \geq \lambda) \\
&= P(Z^2 \geq 0.1979) \\
&\approx 0.7
\end{aligned}
$$

So there is no evidence against H_0.

24 July 5, 2017

24.1 Simple Linear Regression Model (SLRM)

Note that we have Y_i = variable of interset or *response variate* (e.g. STAT231 score of student i).

We also have x_i as the *explanatory variate* used to explain the variate (e.g. STAT230 score of student i).

We will try to estimate the relationship between x_i and Y_i using sample of observation.

Some assumptions we make

(i) Y_i are all independent

(ii) Y_is are normally distributed given x_i. If we graphed all $y_i \in Y_i$s for a given $x \in x_i$ value, then we will see that the Y_i is normally distributed.

(iii) $E(Y_i) = \alpha + \beta x_i$ (or the mean of the Y_is is a linear function of each x_i) where α, β are unknown constants (take the expectation of both sides; $E(R_i) = 0$)

(iv) $Var(Y_i) = \sigma^2$ for all i That is we assume σ^2 does not depend on X (*homoscedastic*; if $\sigma^2 = \sigma^2(x)$ then it will be a *heteroscedastic* model) This is perilous since oftentimes this is not the case.

These 4 assumptions are called the **Gauss-Markov** assumptions.

With these assumptions, we can construct the model

$$Y_i \sim G(\alpha + \beta x_i, \sigma)$$

where Y_i independent, which is equivalent to

$$Y_i = \alpha + \beta x_i + R_i$$

where $R_i \sim G(0, \sigma)$ ($\alpha + \beta x_i$ is the systematic part). The degrees of freedom is $n - \#$ of unknowns in systematic $= n - 2$.

24.2 Finding the SLRM model using MLEs

Given a sample $\{(x_1, y_1), \dots, (x_n, y_n)\}$ what are the MLE for $\hat{\alpha}, \hat{\beta}, \hat{\sigma}$?

Note that the density of Y is

$$f(y) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(y-(\alpha+\beta x))^2}$$

The likelihood function is

$$L(\alpha, \beta, \sigma) = \frac{1}{(2\pi)^{\frac{n}{2}}\sigma^n} e^{-\frac{1}{2\sigma^2}\sum(y_i-(\alpha+\beta x_i))^2}$$

So the log-likelihood function is

$$l(\alpha, \beta, \sigma) = -\frac{n}{2}ln(2\pi) - nln\sigma - \frac{1}{2\sigma^2}\sum(y_i - (\alpha + \beta x_i))^2$$

We must take the FOC $\frac{\partial L}{\partial \alpha} = 0$, $\frac{\partial L}{\partial \beta} = 0$, $\frac{\partial L}{\partial \sigma} = 0$. Solving the equations, we get

$$\begin{aligned}
\hat{\alpha} &= \bar{y} - \hat{\beta}\bar{x} \\
\hat{\beta} &= \frac{S_{xy}}{S_{xx}} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} \\
\hat{\sigma^2} &= \frac{1}{n}\sum(y_i - (\hat{\alpha} + \hat{\beta}x_i))^2 = \frac{S_{yy} - \hat{\beta}S_{xy}}{n}
\end{aligned}$$

where $S_{yy} = \sum(y_i - \bar{y})^2$.

Let's start with $\hat{\beta}$:

1.

$$\begin{aligned}
\hat{\beta} &= \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{S_{xx}} \\
&= \frac{\sum(x_i - \bar{x})y_i}{S_{xx}} \\
&= \frac{\sum x_i(y_i - \bar{y})}{S_{xx}}
\end{aligned}$$

This comes from the fact that the numerator can be manipulated as:

$$\begin{aligned}
\sum(x_i - \bar{x})(y_i - \bar{y}) &= \sum(x_i - \bar{x})y_i - \sum(x_i - \bar{x})\bar{y} \\
&= \sum(x_i - \bar{x})y_i - \bar{y}(\sum x_i - \sum \bar{x}) \\
&= \sum(x_i - \bar{x})y_i - \bar{y}(n\bar{x} - n\bar{x}) \\
&= \sum(x_i - \bar{x})y_i
\end{aligned}$$

24.3 $r_x y$ and $\hat{\beta}$

Does $r_{xy} = 0 \iff \hat{\beta} = 0$?

$$r_{xy} = \frac{S_{xy}}{\sqrt{S_{xx}}\sqrt{S_{yy}}} = \frac{S_{xy}}{S_{xx}} \cdot \frac{\sqrt{S_{xx}}}{\sqrt{S_{yy}}}$$

thus $r_{xy} = \hat{\beta} \cdot \frac{\sqrt{S_{xx}}}{\sqrt{S_{yy}}}$.

24.4 Finding the model using Least Squares

Note we define the square error as

$$\sum e_i^2 = \sum (\hat{y_i} - y_i)^2$$

where $\hat{y_i}$ is the prediction and y_i is the actual.

Choose $\hat{\alpha}, \hat{\beta}$ to minimize

$$\sum_{i=1}^{n} (y_i - (\hat{\alpha} + \hat{\beta} x_i))^2$$

Note these $\hat{\alpha}, \hat{\beta}$ are teh same as the MLEs. They are called the **least square estimates**.

24.5 Mean on Regression Line

Is $(\bar{x}, \bar{y})$ on the regression line?

25 July 7, 2017

25.1 Interpretation of α and β

We have

$$E(Y_i) = \alpha + \beta x_i$$

If we increase x by 1 unit, the average value of Y goes up by β units. *alpha* is the average value of Y when $x = 0$.

25.2 MLE for Sample Variance s^2 (Standard Error)

Remember from before we had the MLE for the variance as

$$\hat{\sigma^2} = \frac{1}{n}(S_{yy} - \hat{\beta} S_{xy})$$

where $S_{yy} = \sum (y_i - \bar{y})^2$.

For our sample variance, instead of subtracting 1 from n, we subtract 2 from n since we have 2 degrees of freedom (α and β in systematic part).

$$\hat{s^2} = \frac{1}{n-2}(S_{yy} - \hat{\beta} S_{xy})$$

Note that s or s_e is also called the **standard error of the regression model**.

25.3 Relationship between $\hat{\beta}$ and Y

Remember that we derived

$$\hat{\beta} = \frac{\sum (x_i - \bar{x}) y_i}{S_{xx}}$$

If we let $a_i = \frac{x_i - \bar{x}}{S_{xx}}$, then we get

$$\hat{\beta} = \sum a_i y_i$$

We can thus even have a random variable (estimator) $\tilde{\beta}$ (distribution of $\hat{\beta}$)

$$\tilde{\beta} = \sum a_i Y_i$$

Since $\tilde{\beta}$ is a linear function of the Y_is, $\tilde{\beta}$ must be Gaussian as well.

25.4 Properties of a_i Expressions

(i)

$$\begin{aligned}
\sum a_i &= \sum \frac{x_i - \bar{x}}{S_{xx}} \\
&= \frac{1}{S_{xx}} \sum (x_i - \bar{x}) \\
&= 0
\end{aligned}$$

Since S_{xx} is a constant and the sum of the deviations from the mean is always 0.

(ii)

$$\begin{aligned}
\sum a_i x_i &= \sum \frac{(x_i - \bar{x}) x_i}{S_{xx}} \\
&= \frac{1}{S_{xx}} \sum (x_i - \bar{x})(x_i - \bar{x}) \\
&= \frac{S_{xx}}{S_{xx}} \\
&= 1
\end{aligned}$$

(iii)

$$
\begin{aligned}
\sum a_i^2 &= \sum \frac{(x_i - \bar{x})^2}{(S_{xx})^2} \\
&= \frac{S_{xx}}{(S_{xx})^2} \\
&= \frac{1}{S_{xx}}
\end{aligned}
$$

25.5 Mean of $\tilde{\beta}$ (Expectation of $\hat{\beta}$)

$$
\begin{aligned}
\tilde{\beta} &= \sum a_i Y_i \\
E(\tilde{\beta}) &= E(\sum a_i Y_i) \\
&= \sum a_i E(Y_i) \\
&= \sum a_i(\alpha + \beta x_i) \\
&= \alpha \sum a_i + \beta \sum a_i x_i \\
&= 0 + \beta \\
&= \beta
\end{aligned}
$$

as desired (Note a_i is not in the expectation since x_is are known).

25.6 Variance of $\tilde{\beta}$

$$
\begin{aligned}
V(\tilde{\beta}) &= V(\sum a_i Y_i) \\
&= \sum a_i^2 V(Y_i) \\
&= \sigma^2 \sum a_i^2 \\
&= \frac{\sigma^2}{S_{xx}}
\end{aligned}
$$

25.7 Distribution of $\tilde{\beta}$

Since $\tilde{\beta}$ is a linear function of Y_i it is Gaussian

$$\tilde{\beta} \sim G(\beta, \frac{\sigma}{\sqrt{S_{xx}}})$$

where the second parameter is the standard deviation.

We can thus form the **pivotal quantity and distribution**

$$\frac{\tilde{\beta} - \beta}{\frac{\sigma}{\sqrt{S_{xx}}}} = Z$$

Note for the sample variance, we can use T_{n-2} (**note the $n-2$!**)

$$\frac{\tilde{\beta}-\beta}{\frac{S}{\sqrt{S_{xx}}}} = T_{n-2}$$

25.8 Confidence Interval for $\tilde{\beta}$

From the t-table (for example a 95% confidence interval)

$$P(-t^* < T_{n-2} < t^*) = 0.95$$
$$P(-t^* < \frac{\tilde{\beta}-\beta}{\frac{S}{\sqrt{S_{xx}}}} < t^*) = 0.95$$

Thus the 95% confidence interval for β is

$$\tilde{\beta} \pm t^* \frac{s}{\sqrt{S_{xx}}}$$

25.9 Hypothesis Testing Example for Correlation

Note a no correlation relationship corresponds to a $\beta = 0$, hence we can hypothesis test for this where

$$H_0 : \beta = 0$$
$$H_1 : \beta \neq 0$$

Our discrepancy measure would be the pivotal quantity

$$D = |\frac{\tilde{\beta}-\beta_0}{\frac{S}{\sqrt{S_{xx}}}}|$$

where

$$d = |\frac{\hat{\beta}-\beta_0}{\frac{s}{\sqrt{S_{xx}}}}|$$

Thus the p-value is

$$\begin{aligned} p &= P(D \geq d) \\ &= P(|T_{n-2}| \geq d) \end{aligned}$$

26 July 10, 2017

26.1 Recap of SLRM Problems

Least Square Line The line of best fit (or least square line) is

$$Y = \hat{\alpha} + \hat{\beta}x$$

where

$$\hat{\alpha} = \bar{y} - \hat{\beta}x$$
$$\hat{\beta} = \frac{S_{xy}}{S_{xx}}$$

these are also the MLEs.

Confidence Interval Remember we have a $\tilde{\beta}$ as the random estimator that corresponds to $\hat{\beta}$. From our previous theorem, we derived pivotal quantities

$$\frac{\tilde{\beta} - \beta}{\frac{S_e}{\sqrt{S_x x}}} \sim T_{n-2}$$

and similarly

$$\frac{(n-2)S_e^2}{\sigma^2} \sim X_{n-2}^2$$

Thus we can derive the confidence interval for a given t^* corresponding to the t-score of a T_{n-2} distribution

$$\hat{\beta} \pm t^* \frac{s_e}{S_{xx}}$$

Hypothesis Testing We can do hypothesis testing with $H_0 : \beta = \beta_0$ and $H_1 : \beta \neq \beta_0$ with the discrepancy measure

$$D = |\frac{\tilde{\beta} - \beta_0}{\frac{S_e}{S_x X}}|$$

and point estimate

$$d = |\frac{\hat{\beta} - \beta_0}{\frac{s_e}{S_x X}}|$$

with p-value

$$P(D \geq d)$$
$$= P(|T_{n-2}| \geq d)$$

Confidence Interval for Mean Response See section below.

Prediction Interval for Y_{new}

26.2 Mean Response (SLRM)

We can $\mu(x) = \alpha + \beta x$ the mean response at a given x value (the average values for all Y for a given x).

Suppose $x = 75$, we want to find the 95% CI for $\mu(75) = \alpha + \beta \cdot 75$ or

$$\hat{\mu} = \hat{\alpha} + 75\hat{\beta}$$

Where $\tilde{\mu}(75)$ is an estimator corresponding to $\hat{\mu}$ or

$$\tilde{\mu} = \tilde{\alpha} + 75\tilde{\beta}$$

and

$$\tilde{\alpha} = \bar{Y} - \tilde{\beta}\bar{x}$$

Note that $\tilde{\alpha}$ is Gaussian since $\tilde{\beta}$ is Gaussian. We can thus derive the Gaussian distribution for $\tilde{\mu}$

$$\tilde{\mu}(x) \sim G(\mu(x), \sigma\sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}})$$

We can thus form the pivotal quantity

$$\frac{\tilde{\mu}(x) - \mu(x)}{\sigma\sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}}} = Z$$

since we want to use the standard error S_e instead

$$\frac{\tilde{\mu}(x) - \mu(x)}{S_e\sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}}} \sim T_{n-2}$$

26.3 Confidence Interval for Mean Response (SLRM)

We thus derive the confidence interval

$$\hat{\mu}(x) \pm t^* s_e\sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}}$$

where $\hat{\mu}(x) = \hat{\alpha} + \hat{\beta}x$. We can find α, β from our sample and x is some arbitrary x value we want to find $\hat{\mu}$ for.

26.4 Confidence Interval for Alpha

Note in the CI for $\mu(x)$, we can plug in $x = 0$ to find the CI for α

$$\hat{\mu}(0) \pm t^* s_e\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}}$$
$$= \alpha \pm t^* s_e\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}}$$

26.5 Confidence Interval for Variance σ^2 (SLRM)

Note we have the pivot

$$\frac{(n-2)S_e^2}{\sigma^2} \sim X_{n-2}^2$$

26.6 Prediction Interval for Y_{new} (SLRM)

Given $x = x_{new}$ (e.g. $x_{new} = 80$), what is the 95% prediction interval for Y_{new}.

Note that the MLE for $Y_{new} = \hat{\alpha} + \hat{\beta} x_{new} = \mu_{new}$. Thus the distribution for Y_{new} is

$$Y_{new} \sim G(\alpha + \beta x_{new}, \sigma)$$

Note we also have for our mean

$$\tilde{\mu}_{new} \sim G(\alpha + \beta x_{new}, \sigma\sqrt{\frac{1}{n} + \frac{(x_{new} - \bar{x})^2}{S_{xx}}})$$

Lets subtract $Y_{new} - \tilde{\mu}_{new}$ (to come up with a pivotal quantity for Y_{new})

$$Y_{new} - \tilde{\mu}_{new} \sim G(0, \sigma\sqrt{1 + \frac{1}{n} + \frac{(x_{new} - \bar{x})^2}{S_{xx}}})$$

Thus we have the pivotal quantity

$$\frac{Y_{new} - \tilde{\mu}_{new}}{S_e\sqrt{1 + \frac{1}{n} + \frac{(x_{new} - \bar{x})^2}{S_{xx}}}} \sim T_{n-2}$$

So our prediction interval for Y_{new} is

$$\hat{\mu}_{new} \pm t^* s_e\sqrt{1 + \frac{1}{n} + \frac{(x_{new} - \bar{x})^2}{S_{xx}}}$$

where $\hat{\mu}_{new} = \hat{\alpha} + \hat{\beta} x_{new}$.

27 July 12, 2017

27.1 Residual

The estimated residual is

$$\begin{aligned} \hat{r_i} &= y_i - \hat{y} \\ &= y_i - (\hat{\alpha} + \hat{\beta} x_i) \end{aligned}$$

These values can be calculated from our sample where y_i are the actual values and $\hat{y}$ is the predicted value. Thus

$$R_i = Y_i - (\alpha + \beta x_i)$$

If the model is correct, $\hat{r_i}$ should act like R_is e.g. $G(0, \sigma)$.

27.2 Standardized Residuals

Note that we can standardize the residuals with respect to the standard error s_e

$$\hat{r_i}^* = \frac{\hat{r_i}}{s_e}$$

Thus

$$\hat{R_i}^* = \frac{\hat{R_i}}{s_e}$$

If the model is correct, then $\hat{r_i}^*$ are outcomes of $G(0, 1)$ random variable.

27.3 Tests for Assumptions

Scatter Plot We plot (x_i, y_i) and look for evidence of linearity.

Residual Plot We can either plot $(x_i, \hat{r_i})$ or $(\hat{\mu_i}, \hat{r_i})$ (and similarly for standardized residuals). We expect plot to form a narrow band around zero. For standardized residuals, $\hat{r_i}$ fall in between $[-3, 3]$ (since 3 corresponds to a 99.7% p-value).

We look for the absence of any pattern. If there is a pattern, then there is evidence of dependency. That is, if the dispersion of $\hat{r_i}$ changes with x then there is evidence of heteroscedasticity.

Q-Q plot Draw Q-Q plot of $\hat{r_i}^*$. If assumptions are true (R_i is normal) then Q-Q plot is a 45 degrees line through the origin.

27.4 Comparing Two Populations

We want to compare two different populations and see whether they are similar in some way. For example, we could test two population from a medical test or test for discrimination between men and women.

Test for Means We have the hypotheses

$$H_0 : \mu_1 = \mu_2$$
$$H_1 : \mu_1 \neq \mu_2$$

where μ_1 and μ_2 are the means of the 1st and 2nd population, respectively.

Test for Proportions We have the hypotheses

$$H_0 : \pi_1 = \pi_2$$
$$H_1 : \pi_1 \neq \pi_2$$

where π_1 and π_2 are the proportions of successes.

28 July 14, 2017

28.1 Hypothesis Testing of Equality of Two Means (Matched Data)

Matched Pair problem: There is a relationship between units of the two populations. Let us have a "before" and "after" population

$$B_1, B_2, \ldots, B_n \sim G(\mu_1, \sigma_1)$$
$$A_1, A_2, \ldots, A_n \sim G(\mu_2, \sigma_2)$$

Where we want to tset $H_0 : \mu_1 = \mu_2, H_1 : \mu_1 \neq \mu_2$. We define

$$Y_i = A_i - B_i$$

which is equivalent to

$$A_i - B_i \sim G(\mu_2 - \mu_1, \sqrt{\sigma_1^2 + \sigma_2^2})$$

where our sample is $\{y_1, \ldots, y_n\}$ where $y_i = a_i - b_i$.

Thus Y_i is Normal

$$Y \sim G(\mu, \sigma)$$

where $\sigma = \sqrt{\sigma_1^2 + \sigma_2^2}$ and $\mu = \mu_2 - \mu_1$ (Gaussian problem with unknown mean and variance).

Our original test thus becomes $H_0 : \mu = 0$

$$D = |\frac{\bar{Y} - 0}{\frac{S}{\sqrt{n}}}|$$

and

$$d = |\frac{\bar{y} - 0}{\frac{s}{\sqrt{n}}}|$$

where $s^2 = \frac{1}{n-1} \sum (y_i - \bar{y})^2$. With the p-value

$$\begin{aligned} p &= P(D \geq d) \\ &= P(|T_{n-1}| \geq d) \end{aligned}$$

We use degrees of freedom of 1 because there is only one unknown, $\bar{y}$ which comes by subtracting each pair of data points. See the unmatched data for an example of $df = 2$.

This same technique can be used for difference of the means e.g. $H_0 : \mu_2 = \mu_1 + 5, H_1 : \mu_2 \neq \mu_1 + 5$.

28.2 Confidence Interval for Difference of Two Means

Can we construct the CI for $\mu_1 - \mu_2$? Yes! Use the Y_i distribution. Left as exercise.

28.3 Unmatched Data

What if there was no underlying natural pairing between two populations?

Our models are

$$Y_{1,i} \sim G(\mu_1, \sigma)$$
$$Y_{2,j} \sim G(\mu_2, \sigma)$$

Assumptions: the variances are equal (need to be supported by the data).

Thus our hypotheses are

$$H_0 : \mu_1 = \mu_2$$
$$H_1 : \mu_1 \neq \mu_2$$

Method 1 We use their sample mean distributions! We have sample mean distributions

$$\bar{Y_1} \sim G(\mu_1, \frac{\sigma}{\sqrt{n_1}})$$
$$\bar{Y_2} \sim G(\mu_2, \frac{\sigma}{\sqrt{n_2}})$$

We take the difference of the two distributions thus we get

$$\bar{Y_1} - \bar{Y_2} \sim G(\mu_1 - \mu_2, \sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}})$$

which is also Gaussian. Our pivotal quantity is

$$\frac{\bar{Y_1} - \bar{Y_2} - (\mu_1 - \mu_2)}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = Z$$

We want to see if $\mu_1 - \mu_2 = 0$ with the test statistic

$$D = |\frac{(\bar{Y_1} - \bar{Y_2}) - 0}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}|$$

We do not know σ though! Let's use the unbiased estimate of σ which is the sample deviation s

$$D = |\frac{(\bar{Y_1} - \bar{Y_2}) - 0}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}|$$

which is similar to $T_{n_1+n_2-2}$ (two unknowns μ_1, μ_2).

For the distribution of our sample variance, we have

$$S^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

Method 2 We perform a regression on the two datasets combined with a binary independent variate. For example for two datasets for income, one for men and one for women, we let men have $x = 0$ and women have $x = 1$. Thus we have a bivariate dataset with $n_1 + n_2$ datapoints.

Note when we try to draw the least square line, the intercept at $x = 0$, $\alpha = E(Y)$ when $x = 0$, is the average men's salary or μ_1.

Note that β is the change of average income when we go from men to women. If $\beta > 0$, women's salary is higher. If $\beta < 0$, men's salary is higher.

Testing for equality of means is testing for $\beta = 0$ in our constructed regression problem, which we've done before!

$$D = |\frac{\tilde{\beta} - 0}{\frac{S_e}{\sqrt{S_{xx}}}}|$$

Our p-value is thus

$$\begin{aligned} p &= P(D \geq d) \\ &= P(|T_{n_1+n_2-2}| \geq d) \end{aligned}$$

This method works only if the variances are equal.

29 July 17, 2017

29.1 Recap of Comparing Distributions

1. Is the data matched? Pair each datapoint such that $Y_i = A_i - B_i$ where Y is our new distribution. Thus our hypothesis becomes $H_0 : \mu_y = 0$.

$$D = |\frac{\bar{Y} - 0}{\frac{S}{\sqrt{n}}}| \sim |T_{n-1}|$$

If no...

2. Are the variances equal? If yes

$$D = |\frac{\bar{Y_1} - \bar{Y_2}}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}| \sim |T_{n_1+n_2-2}|$$

where

$$S^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

If no...

3. Are the sample sizes large? $(n_1 \geq 30, n_2 \geq 30)$ If yes

$$D = |\frac{\bar{Y_1} - \bar{Y_2}}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}| \sim |Z|$$

   Note it is Z and not T since for large $n_i \geq 0$, T-values $\approx$ Z-values.

   If the sample sizes are not large, ignore.

For paired data, the paired test is "more powerful" than the unpaired test.

29.2 Vector Parameters

In some cases, the unknown parameter of interest is a vector

$$\theta = (\theta_1, \ldots, \theta_m)$$
$$H_0 : \theta = \theta(d)$$

(with some restrictions on the values of θ).

Note that the Likelihood Ratio Test Statistic was previously defined for a scalar θ, where $\Lambda(\theta) = X_1^2$.

Theorem. If θ is a vector, then the Likelihood Test Statistic follows

$$\Lambda(\theta) = -2log\frac{L(\theta)}{L(\tilde{\theta})} \sim X_n^2$$

where n is the *number of independent, unrestricted parameters of* θ or the number of parameters estimated under H_0.

Example. A die is rolled. Test whether the die is fair.

Roll the die 300 times, and we count the # of 1s, 2s, etc.

Let θ_i be the probability of rolling i. Then our hypotheses are

$$H_0 : \theta_1 = \theta_2 = \ldots = \theta_6 = \frac{1}{6}$$
$$H_1 : \text{ otherwise}$$

Note our $n = df = 5$ since we have 5 choices/degrees of freedom (the 6th is simply 1 subtract the rest since $\sum \theta_i = 1$).

If we instead tested for only $i = 1, 6$ where

$$H_0 : \theta_1 = \theta_6 = \frac{1}{6}$$
$$H_1 : \text{ otherwise}$$

then we have $df = 5 - 3 = 2$.

29.3 Degrees of Freedom

Thus the degrees of freedom for n categories and k parameters estimated under H_0 is

$$df = (n-1) - k$$

where $n-1$ is equivalent to the # of free choices.

Typically we want a higher degree of freedom (less estimated parameters, more categories).

29.4 Likelihood Test Statistic for Multinomial

So for a Multinomial problem,

$$\Lambda = 2\sum Y_i ln\frac{Y_i}{E_i} \sim X^2_{n-1-k}$$

where

$$\begin{aligned} Y_i &= \text{ observed frequency of category } i \\ E_i &= \text{ expected frequency of category } i \text{ if } H_0 \text{ is true} \end{aligned}$$

That is $E_i = n \times p_i$ where p_i is the expected probability of category i under H_0, n the sample size.

Note the closer the observed to the expected frequency is for each class, then the smaller and better our statistic is.

29.5 Goodness of Fit - Multinomial

Example. In $n = 300$ die rolls, the e_i for each $i \in \{1, \ldots, 6\}$ is $300(\frac{1}{6}) = 50$.

The Likelihood Test Statistics gives us a measure of how close our observed frequencies align with our expected frequencies. For observed frequencies $y = (40, 60, 55, 45, 50, 50)$, we have

$$\begin{aligned} \lambda &= 2\sum y_i ln\frac{y_i}{e_i} \\ &= 2(40ln\frac{40}{50} + 60ln\frac{60}{50} + \ldots) \end{aligned}$$

So for our p-value we have

$$\begin{aligned} p &= P(\Lambda \geq \lambda) \\ &= P(X^2_5 \geq \lambda) \end{aligned}$$

and draw the appropriate conclusions.

30 July 19, 2017

30.1 Goodness of Fit - Arbitrary Frequencies

Example. Four people $1, 2, 3, 4$ are playing poker. Let θ_i be the probability player i wins the poker game.

Let $H_0 : \theta_1 = \theta_2 = 0.4, \theta_3 = \theta_4 = 0.1$.

We test H_0 with an appropriate sample. A sample of 200 days are taken, where we get the observations as follows. We also construct the expected frequency assuming H_0 is true.

Player	y_i = number of wins	e_i
1	90	$200 \times 0.4 = 80$
2	70	$200 \times 0.4 = 80$
3	25	$200 \times 0.1 = 20$
4	15	$200 \times 0.1 = 20$

We calculate $\lambda(\theta)$

$$\begin{aligned}\lambda(\theta) &= 2\sum y_i ln\frac{y_i}{e_i}\\ &= 2(90ln\frac{90}{80} + 70ln\frac{70}{80} + \ldots)\end{aligned}$$

We then calculate the p-value. Note $df = (4-1) - 0 = 3$.

$$\begin{aligned}p &= P(\Lambda \geq \lambda)\\ &= P(X_3^2 \geq \lambda)\end{aligned}$$

30.2 Goodness of Fit - Poisson

Example. Let $X_1, \ldots, X_n$ be samples where $H_0 : X_i \sim Pois(\theta)$.

We have the following frequencies

Values of x	y_i	e_i
0	10	$n \times \hat{p_0}$
1	25	$n \times \hat{p_1}$
2	15	$n \times \hat{p_2}$
3	10	$\vdots$
≥ 4	10	

Let us assume it is poisson, that is $\hat{\mu} = \bar{x}$. We want to first calculate the probabilities of each class:

$$\begin{aligned}P(X=0) = \hat{p_0} &= \frac{e^{-\hat{\mu}}\hat{\mu}^0}{0!}\\ P(X=1) = \hat{p_1} &= \frac{e^{-\hat{\mu}}\hat{\mu}^1}{1!}\\ &\vdots\end{aligned}$$

where $e_i = n \times \hat{p_i}$.

We caclulate the test statistic

$$\lambda(\theta) = 2 \sum y_i ln \frac{y_i}{e_i}$$
$$= \text{known value}$$

With this we can calculate the p-value using $df = (5-1)-1 = 3$. Note we subtracted 1 for number of estimated parameters since we esimated μ.

If instead we had $H_0 : Pois(3)$, then $df = (5-1)-0 = 4$.

30.3 Goodness of Fit - Intervals (for Continuous Data) and Exponential

Example. $X_1, \dots, X_n$ independent r.v.s and $H_0 : X_i \sim Exp(\theta)$.

We have a sample of $n = 200$ with $\bar{x} = 25$.

The samples for $x = ([0,10],[10,20],[20,40],\geq 40)$ are $y = (20,70,80,30)$.

Note we have intervals for our x values, thus we need to take the integral for $\hat{p_i}$. For example (where $\hat{\theta} = \bar{x}$

$$\hat{p_1} = P(x \in [0,10])$$
$$= \int_0^{10} \frac{1}{\hat{\theta}} e^{-\frac{x}{\hat{\theta}}} dx \qquad = \int_0^{10} \frac{1}{\bar{x}} e^{-\frac{x}{\bar{x}}} dx$$

We can then calculate λ the p-value $P(\Lambda \geq \lambda)$ where $df = (4-1)-1 = 2$ (estimated θ) or $P(X_2^2 \geq lambda)$.

31 July 21, 2017

31.1 Restrictions on Multinomial LRTS with Intervals

Note in order for our test statistic to hold, we need

1. n needs to be large ($n \geq 50$)
2. $y_i \geq 5 \forall i$

If we given intervals with $y_i < 5$ e.g. $y_{[7,12)} = 4$, $y_{[12,16]} = 3$, we can collapse these intervals together into $y_{[7,16]} = 7$.

How the data is divided into different categories might affect the final analysis (e.g. 3 categories vs 5 categories). This is difficult to determine.

31.2 Goodness of Fit - Normal

Note if $H_0 : X_i \sim G(\mu, \sigma)$, then k in df is 2 (since we need to estimate both μ and σ). If σ is given then we only estimate $\hat{\mu}$ thus $k = 1$.

31.3 Recap on Goodness of Fit Tests

1. Divide the data into categories and compute the frequencies (y_i)
2. Estimate θ ($\hat{\theta}$ MLE) and use $\hat{\theta}$ to estimate $\hat{p}_i$ assuming H_0 true (and thus e_i)
3. Use the LRTS for multinomial data to find λ, df for X^2_{df}, and p-value

31.4 Test for Independence for Categorical Variates (Contingency Tables)

Are the two variables C going to college and T being a Trump supporter dependent?

We collect a sample and construct a **contingency table**.

	T	not T	
C	20	60	80
not C	40	80	120
	60	140	200

Note if they are independent (H_0), then

$$\begin{aligned} P(C \cap T) &= P(C) \cdot P(T) \\ &= \frac{80}{200} \cdot \frac{60}{200} \end{aligned}$$

So for our expected frequency for C and T to occur at the same time is

$$\begin{aligned} e_{CT} &= n \times \hat{p}_{CT} \\ &= 200 \times \frac{80}{200} \cdot \frac{60}{200} \\ &= \frac{80 \times 60}{200} \end{aligned}$$

In general, for a given row i and column j, the expected frequency e_{ij} is

$$e_{ij} = \frac{r_i \times c_j}{n}$$

where r_i is the sum of row i and c_j is the sum of column j.

We can thus construct the test statistic

$$\lambda = 2 \sum_i \sum_j y_{ij} ln \frac{y_{ij}}{e_{ij}}$$

Our hypothesis is typically

$$H_0 : \theta_{ij} = \alpha_i \beta_j, \forall i, j$$

Our k value is 2 because once we know a θ_{ij} for a fixed row i, then we know the other θ_{ik} for that fixed row i (we can subtract the total number of people in that category to find θ_{ik}). That is we need to know at least θ per row and per column. Thus degrees of freedom is $(4-1)-2=1$.

In general our degrees of freedom for a rows and b columns is

$$\begin{aligned} df &= (n-1)-k \\ &= (ab-1)-((a-1)+(b-1)) \\ &= (a-1)(b-1) \end{aligned}$$

31.5 Tests for Equality of Proportions

We want to test if

$$H_0 : \pi_1 = \pi_2$$

where π_1 is the proportion of smokers with a college degree and π_2 are smokes without a college degree.

	Smoker	Non-Smoker
College	y_{11}	y_{12}
No College	y_{21}	y_{22}

Note if they are equal, this is analagous to not being able to tell if someone has a college degree based on the fact that they are a smoker (or not a smoker).

This implies that we are testing if smoking and college are independent (indepedence test).

32 July 24, 2017

32.1 Equal Proportions Example (Independence)

Example. We want to test if the proportions of smokers are equal across incomes groups. If they are, then this implies that smoking is independent of income.

We have *poor* and *rich* for income categories and smoker and non-smoker for smoking categories.

- 64 of rich people smoke
- 240 rich people
- 86 poor people smoke
- 236 poor people

Step 1 Set-up contingency table of observed frequencies

	Smoker	Non-Smoker	
Rich	64	176	240
Poor	86	150	236
	150	326	476

where y_{ij} is the observed frequency of the i,jth group.

Step 2 Construct table of expected frequencies. Note under $H_0 : \pi_1 = \pi_2$ (where π_1 are proportion of rich smokers, π_2 proportion of poor smokers), we have

$$e_{ij} = \frac{r_i \times c_j}{n}$$

For example $e_{11} = \frac{240 \times 150}{476} = 75.6$.

	Smoker	Non-Smoker	
Rich	75.6	164.4	240
Poor	74.4	161.6	236
	150	326	476

Step 3 Calculate λ

$$\begin{aligned}\lambda &= 2 \sum_i \sum_j y_{ij} ln \frac{y_{ij}}{e_{ij}} \\ &= 5.25\end{aligned}$$

Step 4 Calculate p-value

$$\begin{aligned}p &= P(\Lambda \geq \lambda) \\ &= P(X^2_{(a-1)(b-1)} \geq \lambda) \\ &= P(X^2_1 \geq \lambda) \\ &= P(Z^2 \geq 5.25) \\ &= p(Z \leq -\sqrt{5.25}) + P(Z \geq \sqrt{5.25}) \\ &= 0.02\end{aligned}$$

so we have evidence against H_0.

32.2 Confidence Interval of Proportions

Example. From the previous example, let's find the 95% CI for rich smokers. The CI is (from the Binomial distribution/CI)

$$\hat{\pi_1} \pm z^* \sqrt{\frac{\hat{\pi_1}(1 - \pi_1)}{n}}$$

where $\hat{\pi_1} = \frac{64}{240}$.

32.3 Notes about Independence/Goodness of Fit Testing

1. Note that an alternative test is the Relative Risk (subective). LRTS is an objective mathematical test.

2. Applications of goodness of fit tests: see *Freakonomics - Levitt*.

32.4 Design of Experiments

Let X be our explanatory variable and Y be a response variable.

How can we design an experiment to check whether X "causes" Y?

Definition. X causes Y, if all other things equal, a change in X results in a change in Y.

Note that **confounding variables** are additional variables that add noise to our analysis. For example, suppose the number of Big Macs eaten are confounding variables to our analysis of smoking and cancer.

To check for causation, we have to control confounding variables:

Blocking We fix the level of the confounding variables when we collect data (e.g. sample units who eat the same number of Big Macs).

Randomization We divide the sample randomly into two groups: control and testing group, expecting the confounding factors cancelling out (that is we evenly distribute our confounding variables across the two groups and see if there is a statistical difference).