STAT 270 - Chapter 6
Inference: Single Sample

July 16, 2012

Statistical inference

- Use the sample to study the population
- Sampled units might be different from the unsampled units $\Rightarrow$ Uncertainty

Mathematical reasoning: general $\Rightarrow$ specific
Statistical inference: specific $\Rightarrow$ general

Main inferential problems:

- Estimation*
- Testing*
- Prediction

This chapter: Random sampling - Single sample

Estimation

Unknown parameters of a distriution, e.g., μ in $Normal(\mu, 1)$
Point estimation: Use the observed data to provide a number for the unknown parameter
Example: $x_1, \ldots, x_n$ random sample from $Normal(\mu, 1)$. $\hat{\mu} =?$

Focus of the course:
Interval estimation:

- An interval (a, b) is provided where a and b are functions of data
- We have some confidence that the interval contains the unknown parameter

Normal

$X_1, \ldots, X_n$ iid $Normal(\mu, \sigma^2)$ where μ is unknown and σ^2 is known (unrealistic).

$$\bar{X} \sim Normal(\mu, \frac{\sigma^2}{n})$$

$$\frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim Normal(0, 1)$$

Therefore,

$$P(-1.96 < \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} < 1.96) = 0.95$$

by rearranging,

$$P(\bar{X} - 1.96\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + 1.96\frac{\sigma}{\sqrt{n}})$$

Note: The interval is random Replace $\bar{X}$ with the observed sample mean $\bar{x}$ to obtain a **95% confidence interval** for μ.

$(1-\alpha)\%$ confidence interval (CI)

$$(\bar{X} - z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}})$$

where $z_{\frac{\alpha}{2}}$ is the $(1-\frac{\alpha}{2})$100-th percentile of the standard normal distribution.

Note: the interval is a function of the **observed statistic**.

Large sample, known σ^2

$X_1, \ldots, X_n$ iid with $E(X_i)$, $var(X_i) = \sigma^2$ where μ is unknown and σ^2 is known and no assumptions are made about the distribution of X_is. By CLT

$$\frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim Normal(0, 1)$$

and therefore the $(1 - \alpha)\%$ confidence interval for μ is given by

$$(\bar{X} - z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}})$$

Large sample, unknown σ^2

$X_1, \ldots, X_n$ iid with $E(X_i)$, $var(X_i) = \sigma^2$ where both μ and σ^2 are unknown and no assumptions are made about the distribution of X_is. Use the sample standard deviation $s = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$ as an estimate for σ, i.e., replace σ by $\hat{\sigma} = s$:
The $(1 - \alpha)\%$ confidence interval is given by

$$(\bar{X} - z_{\frac{\alpha}{2}} \frac{s}{\sqrt{n}}, \bar{X} + z_{\frac{\alpha}{2}} \frac{s}{\sqrt{n}})$$

Example 6.1

Suppose $X_1, \ldots, X_n$ are heat measurments in degrees Celsius where $n = 100$, $E(X_i) = \mu$ and $var(X_i) = 16$.
(a) If $\bar{x} = 6.1$ construct a 90% confidence interval for μ.
(b) How large should n be such that a 90% CI is no wider than 0.6 degrees Celsius?

Statistical design: Use statistical theory to address questions regarding how to conduct the experiment **before** collecting the data.

Finite sample, Normal, unknown variance

$X_1, \ldots, X_n$ iid $Normal(\mu, \sigma^2)$ where μ and σ^2 are unknown. Use $\hat{\sigma} = s$ as the estimate of σ. We have

$$\frac{\bar{X} - \mu}{\frac{s}{\sqrt{n}}} \sim t_{(n-1)}$$

Student t distribution with $n - 1$ degrees of freedom A $(1 - \alpha)\%$ CI for μ is given by

$$(\bar{x} - t_{n-1,\frac{\alpha}{2}} \frac{s}{\sqrt{n}}, \bar{x} + t_{n-1,\frac{\alpha}{2}} \frac{s}{\sqrt{n}})$$

where $t_{n-1,\frac{\alpha}{2}}$ is the $(1 - \frac{\alpha}{2})100$-th percentile of the t_{n-1} distribution.

t distribution

If $X \sim t_{n-1}$ the pdf of X is given by

$$f(x) = \frac{\Gamma(\frac{n}{2})}{\Gamma(\frac{n-1}{2})\sqrt{\pi(n-1)}}(1 + \frac{x^2}{n-1})^{-\frac{n}{2}} \qquad -\infty < x < \infty$$

- Symmetric, longer tails than the normal pdf
- Probabilities are obtained from table B.1.
- $t_n \to Normal(0,1)$ as $n \to \infty$.

Pivotal quantity: A statistic whose distribution does not depend on the unknown parameters.
e.g.,

$$\frac{\bar{X} - \mu}{\frac{s}{\sqrt{n}}}$$

Interpretation of confidence intervals

(a, b) is a $(1-\alpha)100\%$ CI for μ:

Wrong: with probability $(1-\alpha)$, $\mu \in (a, b)$.
Because μ is the true value fo the parameter which is assumed to be fixed.

Correct interpretation:
Using frequency definition of probability: As we repeat sampling and construct CI's for the generated samples, we expect $(1-\alpha)100\%$ of these CI's contain μ.

Some notes on CI's

- As n gets large the width of the CI decreases:
  more information $\rightarrow$ more precise estimation
- With fixed n as our confidence $(1-\alpha)$ increases, $z_{\frac{\alpha}{2}}$ increases and therefore the width of the CI increases: A wider CI covers a larger part of the parameter space which results in more confidence that it contains the true value of the parameter.
- Confidence intervals are not unique:
  e.g. $(\bar{x} - z_{.04}\frac{\sigma}{\sqrt{n}}, \bar{x} + z_{.01}\frac{\sigma}{\sqrt{n}})$ is an asymmetric 95% CI.
- Symmetric CI's are the shortest.

Binomial case

Suppose $X \sim binomial(n, p)$ where n is known and p is unknown. Suppose $np \geq 5$ and $n(1-p) \geq 5$ so that we can apply the normal approximation

$$X \sim Normal(np, np(1-p))$$

then

$$\hat{p} \sim Normal(p, \frac{p(1-p)}{n})$$

where $\hat{p} = \frac{X}{n}$ is the proportion of the successes.
Then an approximate $(1-\alpha)100\%$ CI for p is given by

$$(\hat{p}_{obs} - z_{\frac{\alpha}{2}}\sqrt{\frac{p(1-p)}{n}}, \hat{p}_{obs} + z_{\frac{\alpha}{2}}\sqrt{\frac{p(1-p)}{n}})$$

where $\hat{p}_{obs} = \frac{x_{obs}}{n}$.

Example

6 marbles out of 15 randomly selected marbles from a bag containing marbles of different colors are red. Construct a 99% CI for the proportion of red marbles in the bag.

Example

Consider the CI $\bar{x}_{obs} \pm z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$.
(a) How much should the sample size n increase to reduce the width by half?
(b) What is the effect of increasing the sample size by a factor of 25?

Hypothesis testing

Addresses scientific questions in the presence of random variation,

Steps:

1. Determine the **null hypothesis** and **alternative hypothesis**:

   H_0: null hypothesis:
   - the statement of no effect
   - assumed to be true at the begining of the testing process
   - the experimenter wishes to reject H_0 using the evidence provided by the data

   H_1: alternative hypothesis:
   - the state that the experimenter attempts to establish by collecting data

   H_0 and H_1 are
   - disjoint
   - the only possible states of nature; exactly one must be true.
   - not interchangeable
2. Collect data
3. Make inference:
   - data compatible with H_0: do not reject H_0
   - data incompatible with H_0: reject H_0

Discussion

Example 6.3
Example 6.4

P-value

Probability of observing a result as extreme or more extreme than what we observed given thet H_0 is true.

small p-value $\Rightarrow$ data incompatible with H_0

Compare p-value with the **significance level** α (.05 if not mentioned): reject H_0 if p-value$< \alpha$.

Example 1

A restaurant's monthly profit has a normal distribution with average \$1500 and standard deviation of \$200. The owner hires a new chef and decides to keep him only of there is a significant increase in the profit. If the profit is \$1650 at the end of the following month will the owner keep or fire the chef?

Read example 6.5.

Example 2 (Example 6.6)

Example 3

It is claimed that in each bag of M&M's chocolate candies there are equal numbers of each color. If we randomly select 15 candies out of a bag and only one of them is yellow, do we believe the claim? (use significance level of $\alpha =$.05)

Examples 4 and 5 (Examples 6.7 and 6.8)

Example 6

Suppose that mice weight has a normal distribution with mean 20 gr and unknown variance. A new nutrition program which is supposed to cause weight loss is tested on 17 mice and the weights are measured. The sample mean and standard deviation are 18 gr and 2 gr respectively. Has the diet been effective? (Use a significance level of .01).

Error probabilities

$$\alpha = P(\text{type I error}) = P(\text{reject } H_0 | H_0 \text{ true})$$

$$\beta = P(\text{type II error}) = P(\text{not reject } H_0 | H_1 \text{ true})$$

	H_0 true	H_1 true
reject H_0	α	$1-\beta$ no error
do not reject H_0	no error	β

Note that:

- A perfect test (no error) does not exist!
- A compromise should be made between α and β

Fix α, let β be a function of the test; controlling α is more important.
Discussion: Example 6.10.

$$1-\beta = \text{power} = P(\text{reject } H_0 | H_1 \text{ is true})$$

Types of hypothesis

Simple hypothesis: Completely specified, e.g., $\mu = \mu_0$

Composite hypothesis: A range of values, e.g., $\mu > \mu_0$

H_1 is usually composite, therefore β and the power $1 - \beta$ are functions of the parameter.

Critical/rejection region: A subset of the ample space where H_0 gets rejected.

Example 7 (Examples 6.11, 6.12 and 6.13)

Statistical significance (p-value$< \alpha$)

Notes:

- Report p-value instead of the final decision based on p-value$< \alpha$.
- $\alpha = 0.05$ is of no magical importance!
- Statistical significance is not necessarily scientific significance: other factors should also be considered.

Example

Consider $X \sim binomial(500, p)$ where we want to test $H_0 : p = .7$ versus $H_1 : p \neq .7$ at $\alpha = .01$.

(a) Find the critical region of the test.

(b) Calculate the power at $p = .6$.