

A Short Review of Basic Statistics

Yu-Tzung Chang and Hsuan-Wei Lee

Department of Political Science, National Taiwan University

2018.09.20

Some quotes

- *“There are three kinds of lies: lies, damned lies, and statistics.”*
– Benjamin Disraeli
- *“Essentially, all models are wrong, but some are useful.”*
– George Box
- *“The only way to find out what will happen when a complex system is distributed is to disturb the system, not merely to observe it passively.”*
– Fred Mosteller and John Tukey

Outline

- Basic terminology in statistics inference
- Basic terminology in probability theory
- Some famous probability distributions
- Sampling distributions
- Central limit theorem
- Estimation and hypothesis testing
- Confidence intervals
- P-value
- Measure of association and test statistics

Parameters, sample statistics, and sampling distribution

- We are interested in a person's "true" weight. In practice, we may have measured weight in a random sample.
- The former quantity is an example of a *parameter* while the latter is the example of a *sample statistic* or *estimator*.
- For a sample of observations (y_1, \dots, y_n) , use
 - sample average $\bar{y} = \frac{\sum_{i=1}^n y_i}{n}$ to estimate the population mean,
 - sample variance $s^2 = \frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n-1}$ to estimate the true variance.

What makes an estimator good?

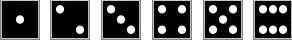
- Two levels of concern:
 - Bias: the difference between this estimator's expected value and the true value of the parameter being estimated. A *systematic error*.
 - Random variance: the variability that is contained within a process that cannot be determined. A *nonsystematic error* (variation between individuals).
- The *sampling distribution* of a statistic is the distribution of that statistic, considered as a random variable, when derived from a random sample of size n . Sampling distributions show how the statistics vary from sample to sample.

Sample space, random variable, and probability function

- The *sample space* of an experiment or random trial is the set of all possible outcomes or results of that experiment.
- A *random variable* is defined as a function that maps the outcomes of unpredictable processes to numerical quantities.
- A *probability function* is a function whose value at any given sample (or point) in the sample space (the set of possible values taken by the random variable) can be interpreted as providing a relative likelihood that the value of the random variable would equal that sample.

Sample space, random variable, and probability function

Example: Roll a die

- sample space: 
- random variable: $\{1, 2, 3, 4, 5, 6\}$
- probability function:
$$Pr(1) = Pr(2) = Pr(3) = Pr(4) = Pr(5) = Pr(6) = \frac{1}{6}$$

Famous discrete probability distribution (*Binomial*(n, p))

- sample space: n independent trials with probability p of success, and probability $(1 - p)$ of failure
- random variable: $Y = \{0, 1, 2, \dots, n\}$, the number of success in n trials
- probability function: $Pr(Y = k) = \binom{n}{k} p^k (1 - p)^{n-k}$

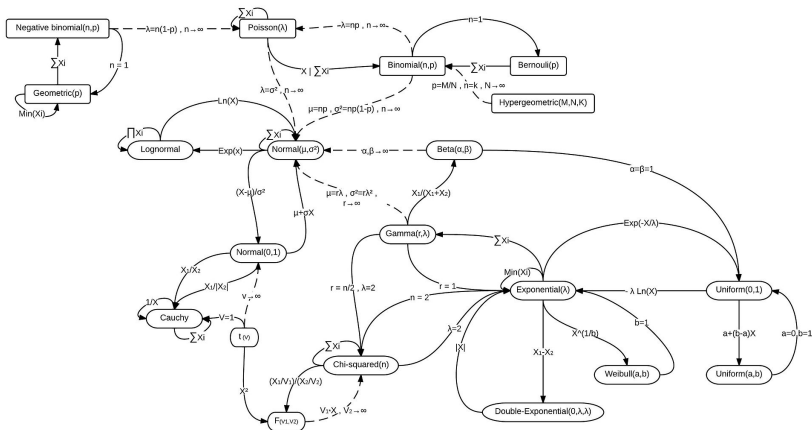
Famous continuous probability distribution ($N(\mu, \sigma)$)

- sample space: target of interest
- random variable: $-\infty < Y < \infty$
- probability function: $f(Y = y) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2\sigma^2}(y - \mu)^2\right]$

Probability distribution

- Probability distribution: relationship between random variable and probability function.
- To determine the associated random variable by using the shape of the probability distribution.
- To determine the percentiles of the distribution by computing the cumulative probability density $Pr(Y \leq y)$.
- To do statistical research based on the characteristics of sampling distributions.

Relationships among probability distributions

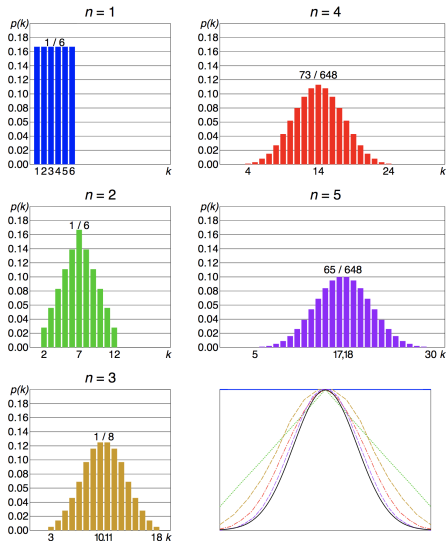


¹LEEMIS, Lawrence M.; Jacquelyn T. MCQUESTON (February 2008).
 "Univariate Distribution Relationships" (PDF). American Statistician. 62 (1):
 45–53.

Sampling distributions

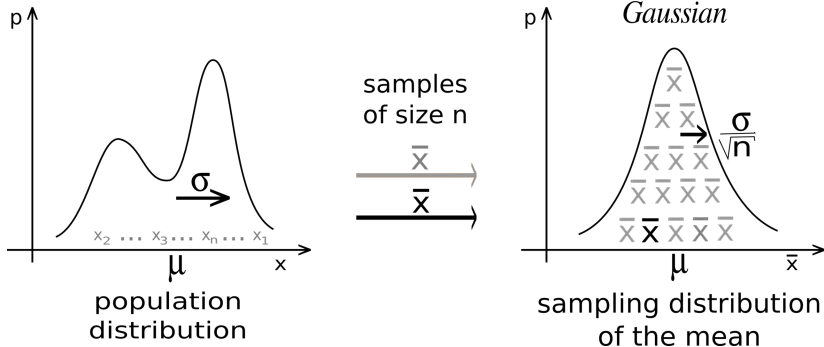
- The probability distribution of a given *random-sample-based statistic* that provides a major simplification of route to *statistical inference*.
- If the sample points follow a normal distribution $N(\mu, \sigma)$, sample means from samples of size n follow a normal distribution $N(\mu, \sigma/n)$. This implies sample means correctly estimate the population mean “on average” and one will have less variation of sample means around the population mean when the sample size is large.
- *Central limit theorem*: Regardless of the distribution of the individual data points, if n is large enough, sample means from sample of size n will be approximately normal, and get closer to normal as n increases.

Central limit theorem (illustration)



¹https://upload.wikimedia.org/wikipedia/commons/8/8c/Dice_sum_central_limit_theorem.svg

Central limit theorem (illustration)



¹Rouaud, Mathieu (2013). Probability, Statistics and Estimation, p. 10.

Central limit theorem (example)

- Toss a unknown biased coin with probability p of getting head ($y_i = 1$) and $1 - p$ of getting tail ($y_i = 0$).
- Toss the coin n times and the sample mean is

$$\sum_{i=1}^n \frac{y_i}{n} = \frac{n_{head} \times 1 + n_{tail} \times 0}{n} = \frac{n_{head}}{n} = p,$$

where p is the proportion of heads in the sample.

- The sampling distribution of a sample proportion is approximately $N(\pi, \pi(1 - \pi)/n)$, where π is the true probability of getting a head.

Two categories of statistical inference (estimation)

- Estimation:
 - *Point estimation*: an estimator of a population parameter, e.g. $\bar{x} \rightarrow \mu$.
 - *Interval estimation*: a point estimate plus an interval that expresses the uncertainty and variability associated with the estimate, e.g. confidence interval.

Two categories of statistical inference (hypothesis testing)

- Hypothesis testing: given the observed data, do we reject or accept a pre-specified null hypothesis in favor of an alternative
- The comparison is deemed statistically significant if the relationship between the data sets would be an unlikely realization of the null hypothesis according to a threshold probability—the significance level.

Confidence interval

- A type of *interval estimate*, computed from the statistics of the observed data, that might contain the true value of an unknown population parameter.
- This quantifies the level of confidence that the parameter lies in the interval.
- Specifically, a confidence interval is defined as an interval that has a specified probability (e.g. 95%) to include the true parameter.
- More strictly speaking, the confidence level represents the *frequency* (i.e. the proportion) of possible confidence intervals that contain the true value of the unknown population parameter.

Confidence interval of population mean

- If the observations (x_1, \dots, x_n) are from $N(\mu, \sigma)$

$$Pr(-1.96 \leq \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \leq 1.96) = 0.95$$

$$Pr(\bar{x} - 1.96 \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + 1.96 \frac{\sigma}{\sqrt{n}}) = 0.95$$

- If the sampling distribution of a statistic is nearly normal and we know (or we can estimate) its standard deviation (standard error), the generic formula that the 95% of confidence interval for any parameter estimated by a statistic that is approximately normally distributed is given by

$$(\text{statistic} \pm 1.96 \times (\text{se of the statistic}))$$

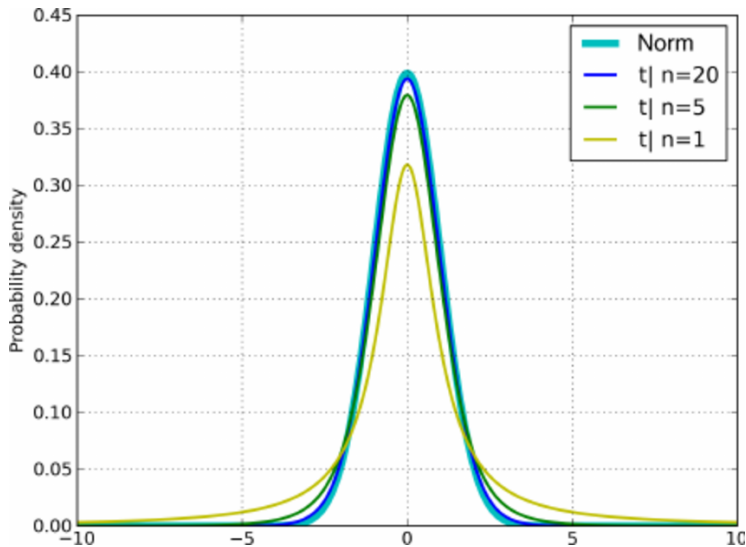
Student's t-distribution and sampling

- So far the discussion has been concentrating on the scenarios where the sampling distribution of a statistic can be assumed to be normal and its variance is known.
- If the sampling distribution of a statistic is normal but its variance σ^2 is unknown, then σ can be estimated by

$$s = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}}$$

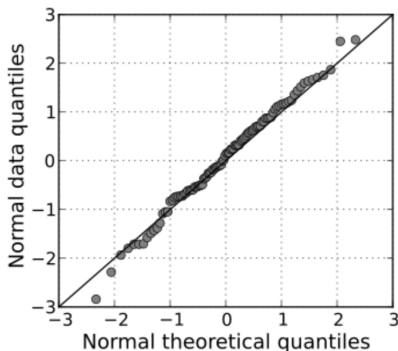
- $\frac{\bar{X} - \mu}{s/\sqrt{n}}$ follows a *t-distribution* with $n - 1$ degrees of freedom
- t-distribution is approximately normal: As n goes large, the approximates the normal distribution

Student's t-distribution and normal distribution



How to tell if a distribution is roughly normal?

- Mean \approx Median \approx Mode?
- Use a histogram to check the skewness
- Quantile-quantile plot (q-q plot)



- Tests for normality, e.g. Kolmogorov-Smirnov test (K-S) and Shapiro-Wilk (S-W) test

Confidence intervals under different assumptions

(X_1, \dots, X_n) are independent random variables with mean μ and variance σ^2 , then the $(1 - \alpha) \times 100\%$ C.I. for μ is

- n : any; normality; σ : known

$$\left(\bar{x} - z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}}, \bar{x} + z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}} \right)$$

- n : any; normality; σ : unknown

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

- $n \gg 30$; normality or non-normality; σ : known or unknown

$$\left(\bar{x} - z_{1-\alpha/2} \frac{s}{\sqrt{n}}, \bar{x} + z_{1-\alpha/2} \frac{s}{\sqrt{n}} \right)$$

Hypothesis testing

One-sample test:

- Compares the mean of the sample to a given number.
- Test whether the population mean weight (μ_{weight}) is equal to 60.
- $H_0 : \mu_{weight} = 60;$
 $H_A : \mu_{weight} \neq 60$

Two-sample test:

- Compare the mean of the first sample minus the mean of the second sample to a given number.
- Test whether the difference of the population mean weights between females and males is equal to 0.
- $H_0 : \mu_f - \mu_m = 0;$
 $H_A : \mu_f - \mu_m \neq 0$

Hypothesis testing

Paired test:

- Compare the mean of the differences in the observations to a given number.
- Test whether the difference of the first measured weight w_1 and the second measured weight w_2 from the *same individual* is equal to 0.
- $H_0 : \mu_{w_1} - \mu_{w_2} = 0;$
 $H_A : \mu_{w_1} - \mu_{w_2} \neq 0$

The distribution of test statistics follow a normal or t-distribution, depending on the information we know about the data.

P-value

- The *p-value* for a hypothesis test is the probability of obtaining a value of the test statistic as extreme or more extreme (in the direction of the critical region) than the one actually computed, when H_0 is true.
- Reporting the p-value associated with a test gives an indication of how common or rare the computed values of the test statistics is given that H_0 is true.
- For judging the significance of a p-value
p-value < 0.05 : “statistical significance” \rightarrow reject H_0
p-value > 0.05 : “non significance” \rightarrow fail to reject H_0

Relationship between commonly used sampling distributions

Normal and χ^2 distribution:

- If Z follows a standard normal $N(0, 1)$ distribution, then Z^2 follows a χ^2 distribution with 1 degree of freedom.
- If $Q = \sum_{i=1}^k Z_i^2$, then $Q \sim \chi^2(k)$.
- Notice that squaring eliminates the sign, so that the square of the 97.5th percentile of the normal is the 95th percentile of the χ^2 , i.e. $1.962^2 = 3.84$. One can think of this as folding over the normal distribution so that both extremes end up at the upper right hand side.

Relationship between commonly used sampling distributions

χ^2 , t and F distributions:

- The F -distribution has two degrees-of-freedom specifications, one for the numerator and the other for the denominator.
- If T follows a t -distribution with ν degree of freedom, then T^2 follows an F -distribution with degree of freedom $(1, \nu)$.
- Let $F_{0.95}(1, 15)$ refer to the 95th percentile of the F -distribution with df $(1, 15)$, and $t_{0.975}$ refer to the 97.5th percentile of the t -distribution with df 15. Then

$$F_{0.95}(1, 15) = 4.541 = 2.131^2 = (t_{0.975}(15))^2.$$

Measure of association

- To compare the difference of certain characteristic between two or more groups. Measure of association between characteristic and grouping.
- Measure of association for continuous data:
 - e.g., whether a weight losing drug does indeed have a beneficial effect?
 - the difference $\mu_1 - \mu_2$ in means between two populations, regression coefficients, or correlation coefficients.
- Measure of association for binary factors:
 - e.g., Is smoking causing the lung cancer?
 - relative risk or odds ratio

The difference of means

- Interested in the difference of means $\mu_1 - \mu_2$.
- The difference is naturally estimated by the difference in sample averages. $\bar{x}_1 - \bar{x}_2$
- Hypothesis test: $H_0 : \mu_1 - \mu_2 = 0$
- Two-sample t-test

Two-sample t-test under different assumptions

$(X_{11}, \dots, X_{1n_1})$ follow a distribution with mean= μ_1 , variance= σ_1^2 .
 $(Y_{21}, \dots, Y_{2n_2})$ follow a distribution with mean= μ_2 , variance= σ_2^2 .
 X_{1i} 's and Y_{2j} 's are independent.

The hypothesis test for $H_0 : \mu_1 - \mu_2 = 0$

1. n_1, n_2 : any; normal; σ_1, σ_2 : known

$$\text{test statistic} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0, 1)$$

2. n_1, n_2 : any; normal; $\sigma_1 = \sigma_2 = \sigma$: unknown

$$\text{test statistic} = \frac{\bar{x}_1 - \bar{x}_2}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t(n_1 + n_2 - 2)$$

$$\text{where } S_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}.$$

Two-sample t-test under different assumptions (continued)

3. $n_1, n_2 \gg 30$; normal or non-normal; σ_1, σ_2 : known or unknown

$$\text{test statistic} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0, 1)$$

Confidence intervals for $\mu_1 - \mu_2$

1. n_1, n_2 : any; normal; σ_1, σ_2 : known

$$(\bar{x}_1 - \bar{x}_2) \pm z_{1-\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

2. n_1, n_2 : any; normal; $\sigma_1 = \sigma_2 = \sigma$: unknown

$$(\bar{x}_1 - \bar{x}_2) \pm t_{1-\alpha/2}(n_1 + n_2 - 2) S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

3. $n_1, n_2 \gg 30$; normal or non-normal; σ_1, σ_2 : known or unknown

$$(\bar{x}_1 - \bar{x}_2) \pm z_{1-\alpha/2} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Measure of association vs. test statistic

- A measure of association
 - a parameter
 - a special parameter that we choose for practical interpretability and purpose
 - does not depend on the sample size, although the success with which we estimate the measure does
- A test statistics
 - target the question whether a certain magnitude of estimated association could have arisen by chance
 - there is nothing of universal interpretability or relevance about a test statistic
 - very much dependent on sample size

The problems of p-values

- If someone people replicates the study with a larger sample size, we expect them to find the same or similar magnitude of association. However, their test statistics will be larger and more significant (and their p-values smaller).
- For the reasons outline, the p-value can not be used as a measure of association.
- Some better way: Use the confidence interval of the difference instead.