TIARA Summer School on Astrostatistics: Statistical Inference

Yen-Chi Chen

Department of Statistics
University of Washington

Summer 2017

Introduction - 1

- What is statistical inference?
- Statistical inference is about how we use data to infer the underlying population that generates our data.
- This process often involves a statistical model.

Introduction - 1

- What is statistical inference?
- Statistical inference is about how we use data to infer the underlying population that generates our data.
- This process often involves a statistical model.
- Given a data (sample), a statistical model is a probability distribution that describes how this data is generated.
- Using the concept of statistical models, we can simply say that statistical inference is how we use the *observed data* to infer some characteristics of the *unobserved (probability) distribution*.

Introduction - 2

- What will the data look like?
- Here is part of the data in the SDSS: It is about 100 galaxy's log stellar mass:

```
11.26 10.76 11.57 11.12 11.25 11.29 11.32 11.46 10.93 11.34  9.12 11.09 11.11
11.62 11.06 11.22 10.94 11.33 10.45 11.79 11.01 11.40 11.38 11.16 11.19 11.47
11.38 11.24 11.05 11.43 11.26 11.15 11.24 11.20 11.55 11.43 11.22 11.36 11.38
11.27 11.04 11.72 11.27 11.16 10.85 11.45 11.37 11.17 11.25 11.10 11.27 11.41
11.15 11.43 11.23 11.61 11.34 11.64 11.53 11.26 11.19 11.20 11.20 11.52 10.49
11.18 11.19 11.52 11.32 11.46 11.03 11.43 11.26 11.13 11.32 11.92 10.94 11.29
11.58 11.11 11.25 11.69 11.28 11.40 11.33 11.44 11.04 11.31 11.22 11.28 11.18
11.60 11.26 11.15 11.17 11.35 11.43 11.26 11.22 11.06
```

- Often our data is just a collection of numbers.
- A statistical model is a distribution that generates these numbers.

Introduction - 3

- In statistics, the values of our data are viewed as random variables $X_1, \cdots, X_n$ (n : sample size) that are IID from a distribution $P(x)$.
- In most cases, we will further assume that such a distribution $P(x)$ has a probability density function $p(x)$.
- The above procedure will often be simply written as

$$X_1, \cdots, X_n \sim P$$

or

$$X_1, \cdots, X_n \sim p.$$

Parameters and Statistics

- Parameters (of interest): numbers or quantities that are features of the population distribution/density.
  - Examples: mean, median, mode, standard deviation (SD), regression coefficients, $\cdots$
  - Parameters are often unknown quantities that we want to know.

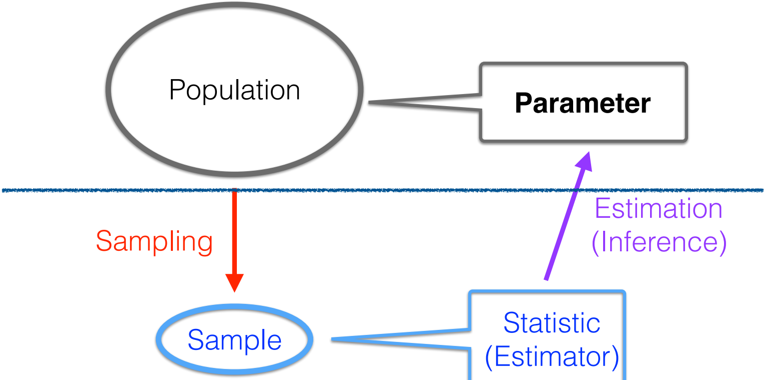
Parameters and Statistics

- Parameters (of interest): numbers or quantities that are features of the population distribution/density.
  - Examples: mean, median, mode, standard deviation (SD), regression coefficients, $\cdots$
  - Parameters are often unknown quantities that we want to know.
- Statistics: numbers or quantities that can be computed using the data.
  - Example: sample average, sample median, sample SD, estimated regression coefficients, $\cdots$
  - From a mathematical point of view, a statistic is a function of random variables.

Parameters and Statistics

- Parameters (of interest): numbers or quantities that are features of the population distribution/density.
  - Examples: mean, median, mode, standard deviation (SD), regression coefficients, $\cdots$
  - Parameters are often unknown quantities that we want to know.
- Statistics: numbers or quantities that can be computed using the data.
  - Example: sample average, sample median, sample SD, estimated regression coefficients, $\cdots$
  - From a mathematical point of view, a statistic is a function of random variables.
- Estimators: when a statistic is used to estimate a parameter, then this statistic is called an estimator (of the corresponding parameter).

Big Picture of Statistical Inference



Estimators and Estimation Theory

Estimating Basic Parameters

- Some parameters have a simple statistic that correspond to each of them.
- If we are interested in estimating these parameters, we can use these simple statistics.
- Example:

$$\text{sample mean} \longleftrightarrow \text{population mean}$$
$$\text{sample median} \longleftrightarrow \text{population median}$$
$$\text{sample SD} \longleftrightarrow \text{population SD}$$

Parametric Model - 1

- In many cases, we assume that the population distribution can be written in certain “parametric form".
- Namely, the density function (or distribution function) is

$$p(x) = p(x; \theta)$$

for some parameter θ.

Parametric Model - 1

- In many cases, we assume that the population distribution can be written in certain "parametric form".
- Namely, the density function (or distribution function) is

$$p(x) = p(x; \theta)$$

  for some parameter θ.
- In parametric model, the parameter θ completely determines the distribution so they are often the parameters of interest.

Parametric Model - 2

Here are some examples of parametric models:

- Normal distribution (parameters: μ and σ):

$$p(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma}}.$$

- Exponential distribution (parameter: λ):

$$p(x; \lambda) = \lambda e^{-\lambda x}, x \geq 0.$$

- Bernoulli distribution (parameter: p):

$$P(X = 1) = p, \quad P(X = 0) = 1 - p.$$

- Poisson distribution (parameter: λ):

$$P(X = k) = \frac{\lambda^k}{k!} e^{-\lambda}, k = 0, 1, 2, \cdots .$$

Maximum Likelihood Estimator - 1

- When we specify a parametric model, we need to estimate the parameter(s) from our data. How to estimate them?

Maximum Likelihood Estimator - 1

- When we specify a parametric model, we need to estimate the parameter(s) from our data. How to estimate them?
- Here is a simple and classical method: **maximum likelihood estimator (MLE)**.
- To illustrate the idea of MLE, we consider the case where we only have one observation. Assume that we observed a number X_1.

Maximum Likelihood Estimator - 1

- When we specify a parametric model, we need to estimate the parameter(s) from our data. How to estimate them?
- Here is a simple and classical method: **maximum likelihood estimator (MLE)**.
- To illustrate the idea of MLE, we consider the case where we only have one observation. Assume that we observed a number X_1.
- For a given parameter θ, according to the parametric model $p(x) = p(x;\theta)$, the probability density generating X_1 is $p(X_i;\theta)$.

Maximum Likelihood Estimator - 1

- When we specify a parametric model, we need to estimate the parameter(s) from our data. How to estimate them?
- Here is a simple and classical method: **maximum likelihood estimator (MLE)**.
- To illustrate the idea of MLE, we consider the case where we only have one observation. Assume that we observed a number X_1.
- For a given parameter θ, according to the parametric model $p(x) = p(x;\theta)$, the probability density generating X_1 is $p(X_i;\theta)$.
- Then we ask a simple question: *for all possible values of the parameter θ, which value has the highest probability density of generating X_1?*

Maximum Likelihood Estimator - 2

- The MLE picks the parameter that has the highest density of generating X_1.
- In other words, we are viewing $p(X_i;\theta)$ as a function of θ and try to find the maximum.

Maximum Likelihood Estimator - 2

- The MLE picks the parameter that has the highest density of generating X_1.
- In other words, we are viewing $p(X_i; \theta)$ as a function of θ and try to find the maximum.
- Thus, we often rewrite it as

$$L(\theta | X_i) = p(X_i; \theta)$$

and such a function is called a *likelihood function*.

Maximum Likelihood Estimator - 2

- The MLE picks the parameter that has the highest density of generating X_1.
- In other words, we are viewing $p(X_i; \theta)$ as a function of θ and try to find the maximum.
- Thus, we often rewrite it as

$$L(\theta|X_i) = p(X_i; \theta)$$

  and such a function is called a *likelihood function*.
- The MLE is then be written as

$$\widehat{\theta}_{MLE} = \underset{\theta}{argmax} \; L(\theta|X_i).$$

Maximum Likelihood Estimator - 3

- In the case of observing n points, $X_1, \cdots, X_n$, the likelihood function is

$$
\begin{aligned}
L(\theta | X_1, \cdots, X_n) &= p(X_1, X_2, \cdots, X_n; \theta) \\
&= p(X_1; \theta) \times \cdots \times p(X_n; \theta).
\end{aligned}
$$

Maximum Likelihood Estimator - 3

- In the case of observing n points, $X_1, \cdots, X_n$, the likelihood function is

$$
\begin{aligned}
L(\theta | X_1, \cdots, X_n) &= p(X_1, X_2, \cdots, X_n; \theta) \\
&= p(X_1; \theta) \times \cdots \times p(X_n; \theta).
\end{aligned}
$$

- The MLE is then be written as

$$\widehat{\theta}_{MLE} = \underset{\theta}{argmax} \;\; L(\theta | X_1, \cdots, X_n).$$

Maximum Likelihood Estimator - 3

- In the case of observing n points, $X_1, \cdots, X_n$, the likelihood function is

$$\begin{aligned} L(\theta|X_1, \cdots, X_n) &= p(X_1, X_2, \cdots, X_n; \theta) \\ &= p(X_1; \theta) \times \cdots \times p(X_n; \theta). \end{aligned}$$

- The MLE is then be written as

$$\widehat{\theta}_{MLE} = \underset{\theta}{argmax} \; L(\theta|X_1, \cdots, X_n).$$

- Note that in the case where the random variables are discrete random variables (such as the Bernoulli model or Poisson model), we will replace the density function by the probability mass function.

Maximum Likelihood Estimator - 4

- In the case of normal distribution, you can find that the MLE $\widehat{\mu}_{MLE}$ and $\widehat{\sigma}^2_{MLE}$ are

$$\widehat{\mu}_{MLE} = \bar{X}_n = \frac{1}{n}\sum_{i=1}^{n}, \quad \widehat{\sigma}^2_{MLE} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2 = \frac{n-1}{n}S_n^2.$$

- The MLE of mean parameter is the sample mean and the MLE of SD parameter is similar to the sample SD.

Maximum Likelihood Estimator - 4

- In the case of normal distribution, you can find that the MLE $\widehat{\mu}_{MLE}$ and $\widehat{\sigma}^2_{MLE}$ are

$$\widehat{\mu}_{MLE} = \bar{X}_n = \frac{1}{n}\sum_{i=1}^{n}, \quad \widehat{\sigma}^2_{MLE} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2 = \frac{n-1}{n}S_n^2.$$

- The MLE of mean parameter is the sample mean and the MLE of SD parameter is similar to the sample SD.
- Note that the parameter p in a Bernoulli distribution and the parameter λ in a Poisson distribution both have the same form of MLE: the sample mean.
- The MLE of an exponential distribution is a bit more interesting.

Maximum Likelihood Estimator - 5

- Recall that an exponential distribution has a probability density function $p(x; \lambda) = \lambda e^{-\lambda x}$.
- Thus, after observing $X_1, \cdots, X_n$, the likelihood function is

$$L(\lambda | X_1, \cdots, X_n) = \prod_{i=1}^{n} \lambda e^{-\lambda X_i} = \lambda^n e^{-\lambda \sum_{i=1}^{n} X_i}.$$

Maximum Likelihood Estimator - 5

- Recall that an exponential distribution has a probability density function $p(x; \lambda) = \lambda e^{-\lambda x}$.
- Thus, after observing $X_1, \cdots, X_n$, the likelihood function is

$$L(\lambda | X_1, \cdots, X_n) = \prod_{i=1}^{n} \lambda e^{-\lambda X_i} = \lambda^n e^{-\lambda \sum_{i=1}^{n} X_i}.$$

- Because taking the logarithm will not affect the position of maximum, we take the log of it. This leads to the log-likelihood function:

$$\ell(\lambda | X_1, \cdots, X_n) = \log L(\lambda | X_1, \cdots, X_n) = n \log \lambda - \lambda \sum_{i=1}^{n} X_i.$$

Maximum Likelihood Estimator - 5

- Recall that an exponential distribution has a probability density function $p(x; \lambda) = \lambda e^{-\lambda x}$.
- Thus, after observing $X_1, \cdots, X_n$, the likelihood function is

$$L(\lambda | X_1, \cdots, X_n) = \prod_{i=1}^{n} \lambda e^{-\lambda X_i} = \lambda^n e^{-\lambda \sum_{i=1}^{n} X_i}.$$

- Because taking the logarithm will not affect the position of maximum, we take the log of it. This leads to the log-likelihood function:

$$\ell(\lambda | X_1, \cdots, X_n) = \log L(\lambda | X_1, \cdots, X_n) = n \log \lambda - \lambda \sum_{i=1}^{n} X_i.$$

- Taking derivative of it and equating it to 0, we obtain

$$\widehat{\lambda}_{MLE} : \frac{n}{\widehat{\lambda}_{MLE}} = \sum_{i=1}^{n} X_i.$$

Maximum Likelihood Estimator - 5

- Recall that an exponential distribution has a probability density function $p(x;\lambda) = \lambda e^{-\lambda x}$.
- Thus, after observing $X_1, \cdots, X_n$, the likelihood function is

$$L(\lambda | X_1, \cdots, X_n) = \prod_{i=1}^{n} \lambda e^{-\lambda X_i} = \lambda^n e^{-\lambda \sum_{i=1}^{n} X_i}.$$

- Because taking the logarithm will not affect the position of maximum, we take the log of it. This leads to the log-likelihood function:

$$\ell(\lambda | X_1, \cdots, X_n) = \log L(\lambda | X_1, \cdots, X_n) = n \log \lambda - \lambda \sum_{i=1}^{n} X_i.$$

- Taking derivative of it and equating it to 0, we obtain

$$\widehat{\lambda}_{MLE} : \frac{n}{\widehat{\lambda}_{MLE}} = \sum_{i=1}^{n} X_i.$$

- Thus, $\widehat{\lambda}_{MLE} = \frac{n}{\sum_{i=1}^{n} X_i} = \frac{1}{\bar{X}_n}$.

Estimation Theory: Bias and Variance - 1

- Let θ be the parameter of interest and $\widehat{\theta}_n$ be an estimator for it.
- How do we quantify the accuracy of the estimator?
- Beware: since the estimator is computed from the data, the randomness of data will propagate to $\widehat{\theta}_n$. So $\widehat{\theta}_n$ is often a random quantity.

Estimation Theory: Bias and Variance - 1

- Let θ be the parameter of interest and $\widehat{\theta}_n$ be an estimator for it.
- How do we quantify the accuracy of the estimator?
- Beware: since the estimator is computed from the data, the randomness of data will propagate to $\widehat{\theta}_n$. So $\widehat{\theta}_n$ is often a random quantity.
- To quantify the accuracy of $\widehat{\theta}_n$, we introduce two measures: bias and variance.
- The bias of an estimator is the systematic deviation from its target. Mathematically, we define it as

$$\mathbf{bias}(\widehat{\theta}_n) = \mathbb{E}(\widehat{\theta}_n) - \theta.$$

Estimation Theory: Bias and Variance - 1

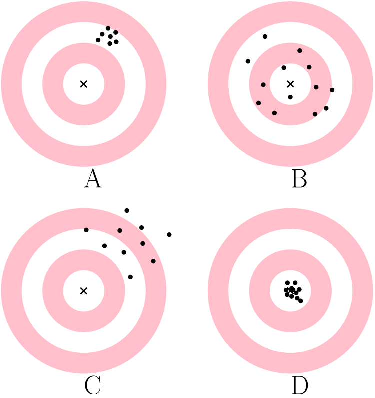
- Let θ be the parameter of interest and $\widehat{\theta}_n$ be an estimator for it.
- How do we quantify the accuracy of the estimator?
- Beware: since the estimator is computed from the data, the randomness of data will propagate to $\widehat{\theta}_n$. So $\widehat{\theta}_n$ is often a random quantity.
- To quantify the accuracy of $\widehat{\theta}_n$, we introduce two measures: bias and variance.
- The bias of an estimator is the systematic deviation from its target. Mathematically, we define it as

$$\textbf{bias}(\widehat{\theta}_n) = \mathbb{E}(\widehat{\theta}_n) - \theta.$$

- The variance describes the amount of randomness that an estimator has. It is simply the quantity $\mathrm{Var}(\widehat{\theta}_n)$, variance of the estimator.

Estimation Theory: Bias and Variance - 2

- A: large bias, small variance.
- B: small bias, large variance.
- C: large bias, large variance.
- D: small bias, small variance.
- Ideally, we want an estimator with small bias and small variance (case D).



Estimation Theory: MSE

- An estimator is called *consistent* if when the sample size $n \to \infty$, $\widehat{\theta}_n$ converges to θ in probability.
- Both the bias and variance converges to 0 $\Longrightarrow$ it is a consistent estimator.

Estimation Theory: MSE

- An estimator is called *consistent* if when the sample size $n \to \infty$, $\widehat{\theta}_n$ converges to θ in probability.
- Both the bias and variance converges to 0 $\Longrightarrow$ it is a consistent estimator.
- There is a simple error measurement that takes into account both the bias and variance called the *mean square error (MSE)*.
- The MSE is

$$MSE(\widehat{\theta}_n, \theta) = \mathbb{E}\left((\widehat{\theta}_n - \theta)^2\right) = \mathbf{bias}^2(\widehat{\theta}_n) + \mathrm{Var}(\widehat{\theta}_n).$$

- The last equality is also known as the *bias-variance decomposition*.

Estimation Theory: Some Remarks

- A good news: most MLE's are consistent, so are the simple estimators of basic parameters.
- The consistency of an estimator depends on the assumptions about the population distribution.
- Even we are using the same estimator, it might be consistent for one dataset but inconsistent for another.
- An inconsistent estimator may lead you to a wrong conclusion.

Confidence Interval

Confidence Interval

- Confidence interval (CI) is an approach that uses an interval to infer the parameter of interest.
- Some people call it *interval estimation* as opposite to the *point estimation* (the regular estimator we just described).

Confidence Interval

- Confidence interval (CI) is an approach that uses an interval to infer the parameter of interest.
- Some people call it *interval estimation* as opposite to the *point estimation* (the regular estimator we just described).
- A CI requires a number called confidence level, often denoted as $1-\alpha$, and a CI is an *interval* $C_{\alpha,n} = [L_{\alpha,n}, U_{\alpha,n}]$ *that can be computed using the data* with the following property:

$$P(\theta \in C_{\alpha,n}) = P(L_{\alpha,n} \leq \theta \leq U_{\alpha,n}) \approx 1-\alpha.$$

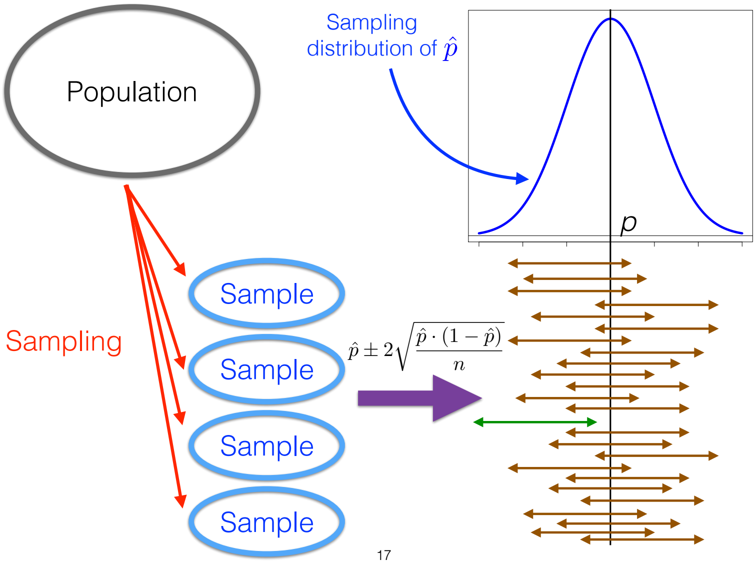
Confidence Interval

- Confidence interval (CI) is an approach that uses an interval to infer the parameter of interest.
- Some people call it *interval estimation* as opposite to the *point estimation* (the regular estimator we just described).
- A CI requires a number called confidence level, often denoted as $1-\alpha$, and a CI is an *interval* $C_{\alpha,n} = [L_{\alpha,n}, U_{\alpha,n}]$ *that can be computed using the data* with the following property:

$$P(\theta \in C_{\alpha,n}) = P(L_{\alpha,n} \leq \theta \leq U_{\alpha,n}) \approx 1-\alpha.$$

- The randomness in the above probability comes from the randomness of $L_{\alpha,n}, U_{\alpha,n}$ not θ!
- The lower bound $L_{\alpha,n}$ and upper bound $U_{\alpha,n}$ are statistics (numbers computed from the data).

Confidence Interval: an Illustration



Confidence Interval: Bernoulli Distribution

- The plot in the previous slide shows an example for a $\approx 95\%$ CI for inferring the parameter p in a Bernoulli distribution.
- The interval with

$$L_n = \widehat{p} - 2\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}, U_n = \widehat{p} + 2\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}, \widehat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

  is a $\approx 95\%$ confidence interval for the parameter p.
- Note that $\widehat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i$ is the MLE for the parameter p.

Confidence Interval: Bernoulli Distribution

- The plot in the previous slide shows an example for a $\approx 95\%$ CI for inferring the parameter p in a Bernoulli distribution.
- The interval with

$$L_n = \widehat{p} - 2\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}, U_n = \widehat{p} + 2\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}, \widehat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

  is a $\approx 95\%$ confidence interval for the parameter p.
- Note that $\widehat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i$ is the MLE for the parameter p.
- Common CIs have three components as indicated by the colors:
  - The estimator.
  - Standard error (SE) of the estimator.
  - Multiplier: determined by $1-\alpha$, the confidence.

Standard Error of an Estimator

- SE is an estimator of the standard deviation of the estimator.
- SE is the rough size of the error of our estimator.
- When reading papers, people often report the estimated value and its error – this error is the SE.

Standard Error of an Estimator

- SE is an estimator of the standard deviation of the estimator.
- SE is the rough size of the error of our estimator.
- When reading papers, people often report the estimated value and its error – this error is the SE.
- For the estimator $\widehat{p}$ in the Bernoulli model, its variance

$$
\begin{aligned}
\mathrm{Var}(\widehat{p}) &= \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}(X_i) \qquad \text{(independence)} \\
&= \frac{1}{n}\mathrm{Var}(X_1) \qquad \text{(identical)} \\
&= \frac{p(1-p)}{n}.
\end{aligned}
$$

- Thus, the SD of $\widehat{p}$ is $\sqrt{\frac{p(1-p)}{n}}$, which can be approximated by $\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}$.

Multiplier - 1

- A multiplier is a number depending on the *distribution of the estimator.*
- Thanks to the central limit theorem, many estimators will be normally distributed around their targeted parameters.
- Namely, if we use $\widehat{\theta}_n$ to estimate θ, under good assumptions we have

$$\widehat{\theta}_n \approx N(\theta, SE^2(\widehat{\theta}_n)).$$

Multiplier - 1

- A multiplier is a number depending on the *distribution of the estimator.*
- Thanks to the central limit theorem, many estimators will be normally distributed around their targeted parameters.
- Namely, if we use $\widehat{\theta}_n$ to estimate θ, under good assumptions we have

$$\widehat{\theta}_n \approx N(\theta, SE^2(\widehat{\theta}_n)).$$

- This implies

$$\frac{\widehat{\theta}_n - \theta}{SE(\widehat{\theta})} \approx N(0, 1).$$

Multiplier - 1

- A multiplier is a number depending on the *distribution of the estimator.*
- Thanks to the central limit theorem, many estimators will be normally distributed around their targeted parameters.
- Namely, if we use $\widehat{\theta}_n$ to estimate θ, under good assumptions we have
$$\widehat{\theta}_n \approx N(\theta, SE^2(\widehat{\theta}_n)).$$
- This implies
$$\frac{\widehat{\theta}_n - \theta}{SE(\widehat{\theta})} \approx N(0,1).$$
- Let $z_{1-\alpha/2}$ be the number such that $P(|N(0,1)| \leq z_{1-\alpha/2}) = 1 - \alpha$.
- Then we have:
$$P\left(\left|\frac{\widehat{\theta}_n - \theta}{SE(\widehat{\theta})}\right| \leq z_{1-\alpha/2}\right) \approx 1 - \alpha.$$

Multiplier - 2

- This fact:
$$P\left(\left|\frac{\widehat{\theta}_n - \theta}{SE(\widehat{\theta})}\right| \le z_{1-\alpha/2}\right) \approx 1-\alpha.$$
implies that
$$P\left(\widehat{\theta}_n - z_{1-\alpha/2}\cdot SE(\widehat{\theta}_n) \le \theta \le \widehat{\theta}_n + z_{1-\alpha/2}\cdot SE(\widehat{\theta}_n)\right) \approx 1-\alpha.$$
- Namely, $\widehat{\theta}_n \pm z_{1-\alpha/2}\cdot SE(\widehat{\theta}_n)$ is a $1-\alpha$ CI of θ.

Multiplier - 2

- This fact:
$$P\left(\left|\frac{\widehat{\theta}_n-\theta}{SE(\widehat{\theta})}\right|\le z_{1-\alpha/2}\right)\approx 1-\alpha.$$
implies that
$$P\left(\widehat{\theta}_n - z_{1-\alpha/2}\cdot SE(\widehat{\theta}_n)\le \theta \le \widehat{\theta}_n + z_{1-\alpha/2}\cdot SE(\widehat{\theta}_n)\right)\approx 1-\alpha.$$
- Namely, $\widehat{\theta}_n \pm z_{1-\alpha/2}\cdot SE(\widehat{\theta}_n)$ is a $1-\alpha$ CI of θ.
- For a normal distribution, $z_{0.975}\approx 1.96 \approx 2$ ($\alpha = 0.05 = 5\%$).
- By identifying $\widehat{\theta}_n = \widehat{p}$ and $SE(\widehat{\theta}_n) = SE(\widehat{p}) \approx \sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}$ and $z_{0.975}\approx 2$, we conclude that $\widehat{p}\pm 2\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}$ is a CI of the parameter p in the Bernoulli distribution.

Confidence Interval: Mean

- Assume that we observe $X_1, \cdots, X_n \sim P$ and we are interested in the population mean μ.
- A simple estimator of the mean is the sample mean $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$.
- The SD of the sample mean is

$$\sqrt{\mathrm{Var}(\bar{X}_n)} = \sqrt{\frac{\sigma^2}{n}},$$

  where σ^2 is the variance of the population distribution (population variance).
- We can simply estimate the population variance by sample variance $S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$.
- Thus, a 95% CI of the mean μ is

$$\left[\bar{X}_n - 1.96 \cdot \frac{S_n}{\sqrt{n}},\ \bar{X}_n + 1.96 \cdot \frac{S_n}{\sqrt{n}}\right].$$

Confidence Interval: Example

- Now going back to the galaxy example mentioned at the beginning. We have $n = 100$ galaxies.
- After computation, we found that the sample mean of log stellar mass $\bar{X}_n = 11.25$ and sample SD is 0.31.
- Thus, the SE of the sample mean is about $\frac{0.31}{\sqrt{100}} = 0.031$
- A 95% CI of the mean log stellar mass is

$$[11.25 - 1.96 \cdot 0.031, \;\; 11.25 + 1.96 \cdot 0.031] = [11.19, 11.31].$$

Confidence Interval: Some Remarks

- The significance level is often chosen by the researcher.
- Common choices are 95%, 90%, and 99% ($\alpha = 0.05, 0.1, 0.01$).
- The corresponding number $z_{1-\alpha/2}$ will be

$$z_{0.975} \approx 1.96,\ z_{0.95} \approx 1.64,\ z_{0.996} \approx 2.58.$$

- Note that a CI can also be one-sided. Namely, it can also be an interval like $(-\infty, c]$ or $[c, \infty)$ for some constant c.
- Often the construction of CIs depends on the (asymptotic/limiting) distribution of the estimator. Normal distribution is a common case but sometimes people will use other distributions such as t-distribution, χ^2-distribution, F-distribution, etc.

Hypothesis Test

Hypothesis Test - 1

- Hypothesis test is a statistical procedure to make inference.
- Actually, many scientific discoveries implicitly or explicitly used the hypothesis test.
- In hypothesis test, we are comparing two hypothesis:
  - Null hypothesis (H_0): a statement we are testing its validity.
  - Alternative hypothesis (H_a): a statement against to the null hypothesis.

Hypothesis Test - 1

- Hypothesis test is a statistical procedure to make inference.
- Actually, many scientific discoveries implicitly or explicitly used the hypothesis test.
- In hypothesis test, we are comparing two hypothesis:
  - Null hypothesis (H_0): a statement we are testing its validity.
  - Alternative hypothesis (H_a): a statement against to the null hypothesis.
- In scientific research, the alternative hypothesis is often something we want to prove (using data) – we will explain this later.

Hypothesis Test - 2

- How do we carry out a test?
- Here is a brief description of the testing procedure.

Hypothesis Test - 2

- How do we carry out a test?
- Here is a brief description of the testing procedure.
    1. We design a test statistic T_n (can be computed from the data).

Hypothesis Test - 2

- How do we carry out a test?
- Here is a brief description of the testing procedure.
  1. We design a test statistic T_n (can be computed from the data).
  2. We study the distribution of such a test statistic *assuming* H_0 *is true*.

Hypothesis Test - 2

- How do we carry out a test?
- Here is a brief description of the testing procedure.
  1. We design a test statistic T_n (can be computed from the data).
  2. We study the distribution of such a test statistic *assuming* H_0 *is true*.
  3. Using the data, we then compute the value of the observed test statistics T_n.

Hypothesis Test - 2

- How do we carry out a test?
- Here is a brief description of the testing procedure.
  1. We design a test statistic T_n (can be computed from the data).
  2. We study the distribution of such a test statistic *assuming* H_0 *is true*.
  3. Using the data, we then compute the value of the observed test statistics T_n.
  4. Using the distribution of test statistic, we compute the *probability of observing an event that is more extreme than our observed test statistic*. This probability is called the P-value.

Hypothesis Test - 2

- How do we carry out a test?
- Here is a brief description of the testing procedure.
  1. We design a test statistic T_n (can be computed from the data).
  2. We study the distribution of such a test statistic *assuming* H_0 *is true*.
  3. Using the data, we then compute the value of the observed test statistics T_n.
  4. Using the distribution of test statistic, we compute the *probability of observing an event that is more extreme than our observed test statistic*. This probability is called the P-value.
  5. We reject H_0 if P-value is smaller than the significance level α.

Hypothesis Test: Example - 1

- Again we use the galaxy stellar mass data as an example.
- Assume that the previous literature suggested that the log stellar mass is
$$H_0 : \mu = 11.15.$$
We want to test if this statement is reasonable using our data. Note that in this case, the alternative hypothesis is
$$H_a : \mu \neq 11.15.$$

Hypothesis Test: Example - 1

- Again we use the galaxy stellar mass data as an example.
- Assume that the previous literature suggested that the log stellar mass is

  $$H_0 : \mu = 11.15.$$

  We want to test if this statement is reasonable using our data. Note that in this case, the alternative hypothesis is

  $$H_a : \mu \neq 11.15.$$

- First we need to choose a test statistic. Here we simply use a rescaled sample average as the test statistic:

  $$T_n = \frac{\bar{X}_n - \mu}{SE} = \frac{\bar{X}_n - 11.15}{SE}.$$

Hypothesis Test: Example - 1

- Again we use the galaxy stellar mass data as an example.
- Assume that the previous literature suggested that the log stellar mass is

  $$H_0 : \mu = 11.15.$$

  We want to test if this statement is reasonable using our data. Note that in this case, the alternative hypothesis is

  $$H_a : \mu \neq 11.15.$$

- First we need to choose a test statistic. Here we simply use a rescaled sample average as the test statistic:

  $$T_n = \frac{\bar{X}_n - \mu}{SE} = \frac{\bar{X}_n - 11.15}{SE}.$$

- Then we need to find the distribution of this test statistic. And it turns out that such a test statistic T_n, under H_0, has a nice distribution:

  $$T_n \approx N(0, 1).$$

Hypothesis Test: Example - 2

- Using the data, we compute the observed value of test statistic:
$$t_n = \frac{11.25 - 11.15}{0.031} = 3.23.$$
Note that sometimes people would interpret this as a signal of 3.23σ.

Hypothesis Test: Example - 2

- Using the data, we compute the observed value of test statistic:
$$t_n = \frac{11.25 - 11.15}{0.031} = 3.23.$$
Note that sometimes people would interpret this as a signal of 3.23σ.
- The P-value is
$$p_{value} = P(|N(0,1)| \geq t_n) \approx 1.24 \times 10^{-3}.$$

Hypothesis Test: Example - 2

- Using the data, we compute the observed value of test statistic:
$$t_n = \frac{11.25 - 11.15}{0.031} = 3.23.$$
Note that sometimes people would interpret this as a signal of 3.23σ.
- The P-value is
$$p_{value} = P(|N(0,1)| \geq t_n) \approx 1.24 \times 10^{-3}.$$
- If we choose a significance level of $5\%, 1\%$, we will all reject H_0. In this case, we will claim that we have strong evidence to reject $H_0 : \mu = 11.25$ under a significance level of 5% (or 1%).

Hypothesis Test: Significance Level

- The significance level can be interpreted as a tolerance level of wrongly reject H_0 (later we will call it type-1 error rate).
- If H_0 is correct, then the distribution of P-value will be a uniform distribution over 0 and 1 (you can try to prove this).

Hypothesis Test: Significance Level

- The significance level can be interpreted as a tolerance level of wrongly reject H_0 (later we will call it type-1 error rate).
- If H_0 is correct, then the distribution of P-value will be a uniform distribution over 0 and 1 (you can try to prove this).
- So when H_0 is correct, the chance of rejecting H_0 under a significance level α is α.
- A lower value of α requires a stronger evidence against H_0 to reject it.

Hypothesis Test: Significance Level

- The significance level can be interpreted as a tolerance level of wrongly reject H_0 (later we will call it type-1 error rate).
- If H_0 is correct, then the distribution of P-value will be a uniform distribution over 0 and 1 (you can try to prove this).
- So when H_0 is correct, the chance of rejecting H_0 under a significance level α is α.
- A lower value of α requires a stronger evidence against H_0 to reject it.
- In our case, we can reject H_0 under $\alpha = 5\%, 1\%$ but not 0.1%.
- Rejecting H_0 implies that the claim $\mu = 11.15$ is not reasonable so we conclude $\mu \neq 11.15$.

Hypothesis Test: the Basic Idea

- What are we doing in hypothesis test?
- Actually, we are doing a generalization of *proof by contradiction*.

Hypothesis Test: the Basic Idea

- What are we doing in hypothesis test?
- Actually, we are doing a generalization of *proof by contradiction*.
- Proof by contradiction: if we want to prove a statement $\mathcal{S}$, we assume that statement to be incorrect first and then show that this leads to a contradiction.
- In the case of hypothesis test, we are doing a very similar job: we first assume H_0 being correct and then show that H_0 contradicts to our data.

Hypothesis Test: the Basic Idea

- What are we doing in hypothesis test?
- Actually, we are doing a generalization of *proof by contradiction*.
- Proof by contradiction: if we want to prove a statement $\mathcal{S}$, we assume that statement to be incorrect first and then show that this leads to a contradiction.
- In the case of hypothesis test, we are doing a very similar job: we first assume H_0 being correct and then show that H_0 contradicts to our data.
- The **P-value** is a quantity that serves as *a measure of consistency between H_0 and our data*. Thus, a low P-value means that H_0 is not consistent with data (i.e., they contradict to each other) so we reject the null hypothesis.
- Thus, you can easily see that the alternative hypothesis will be something we want to prove (because it is the complement of the null hypothesis).

Hypothesis Test: Type-1 and Type-2 Error

- In hypothesis testing, there are two types of errors.
- *Type-1 error*: the H_0 is correct but we mistakenly reject H_0.
- *Type-2 error*: the H_0 is incorrect but we do not reject H_0.

Hypothesis Test: Type-1 and Type-2 Error

- In hypothesis testing, there are two types of errors.
- *Type-1 error*: the H_0 is correct but we mistakenly reject H_0.
- *Type-2 error*: the H_0 is incorrect but we do not reject H_0.
- If we reject H_0 under significance level α, this implies that the type-1 error rate is less than or equal to α.
- This implies: we are controlling type-1 error rate to be small in hypothesis test framework.

Hypothesis Test: Type-1 and Type-2 Error

- In hypothesis testing, there are two types of errors.
- *Type-1 error*: the H_0 is correct but we mistakenly reject H_0.
- *Type-2 error*: the H_0 is incorrect but we do not reject H_0.
- If we reject H_0 under significance level α, this implies that the type-1 error rate is less than or equal to α.
- This implies: we are controlling type-1 error rate to be small in hypothesis test framework.
- Why do we want to control type-1 error rate? $\rightarrow$ We are doing a 'proof by contradiction' so we want to make sure *we have strong evidence to 'prove' that H_0 is incorrect*.
- Rejecting under a very small α. $\Leftrightarrow$ Type-1 error rate is very small. $\Leftrightarrow$ We have very strong evidence.

Hypothesis Test: Other Examples

- Sometimes the null hypothesis will be a one-sided case such as

$$H_0 : \mu \leq 11.15.$$

  In this case, the calculation of P-value will be slightly different since the 'more extreme' case will be on the other side.
- When we assume the parametric model is correct, many null hypothesis will be about the value of parameter. For instance, $X_1, \cdots, X_m \sim N(0, \sigma^2)$ and

$$H_0 : \sigma^2 = 0.2, \quad H_a : \sigma^2 \neq 0.2.$$

- The hypothesis test and CI have a close relationship. In some cases, you can use one to compute the other.

Two-Sample Test - 1

- Two-sample test is a very important topic in scientific research.
- The goal of a two-sample test is to see if we have strong evidence that the two observed samples are from different populations.
- For instance, in analyzing galaxies, we can separate galaxies into two groups: elliptical galaxies and spiral galaxies. The two-sample test can be used to check if the distributions of stellar mass of the two populations are the same or not.

Two-Sample Test - 2

- Let $X_1, \cdots, X_n$ and $Y_1, \cdots, Y_m$ be the two samples we observed.
- Using statistical models, we model that

$$X_1, \cdots, X_n \sim P_X, \qquad Y_1, \cdots, Y_m \sim P_Y,$$

  where P_X and P_Y are the distributions generating the two samples.
- The two-sample test examines the following hypothesis:

$$H_0 : P_X = P_Y$$

  against

$$H_a : P_X \neq P_Y.$$

Two-Sample Test: Mean Test - 1

- A simple approach of the two-sample test is the mean test.
- Because
$$H_0 : P_X = P_Y$$
implies $\mu_X = \mu_Y$ (μ_i is the mean of P_i), the mean test is to test
$$H_0 : \mu_X = \mu_Y.$$
- Testing $\mu_X = \mu_Y$ is equivalent to testing
$$H_0 : \mu_X - \mu_Y = 0.$$
- So the test statistics is to use the difference between sample means $\bar{X}_n$ and $\bar{Y}_m$ and rescale it by the variance.

Two-Sample Test: Mean Test - 2

- The sample means have variance

$$\mathrm{Var}(\bar{X}_n) = \frac{\sigma_X^2}{n}, \quad \mathrm{Var}(\bar{Y}_m) = \frac{\sigma_Y^2}{m},$$

  where σ_X^2 and σ_Y^2 are the variance of P_X and P_Y.
- Thus, the quantity $\bar{X}_n - \bar{Y}_m$ has variance $\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}$ (because they are independent).

Two-Sample Test: Mean Test - 2

- The sample means have variance

$$\mathrm{Var}(\bar{X}_n) = \frac{\sigma_X^2}{n}, \quad \mathrm{Var}(\bar{Y}_m) = \frac{\sigma_Y^2}{m},$$

  where σ_X^2 and σ_Y^2 are the variance of P_X and P_Y.
- Thus, the quantity $\bar{X}_n - \bar{Y}_m$ has variance $\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}$ (because they are independent).
- Because we do not know σ_X^2 and σ_Y^2 in practice, we will replace them by the sample variance S_X^2 and S_Y^2.

Two-Sample Test: Mean Test - 2

- The sample means have variance
$$\mathrm{Var}(\bar{X}_n) = \frac{\sigma_X^2}{n}, \quad \mathrm{Var}(\bar{Y}_m) = \frac{\sigma_Y^2}{m},$$
where σ_X^2 and σ_Y^2 are the variance of P_X and P_Y.
- Thus, the quantity $\bar{X}_n - \bar{Y}_m$ has variance $\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}$ (because they are independent).
- Because we do not know σ_X^2 and σ_Y^2 in practice, we will replace them by the sample variance S_X^2 and S_Y^2.
- Thus, our final test statistics is
$$T_{n,m} = \frac{\bar{X}_n - \bar{Y}_m}{\sqrt{\frac{S_X^2}{n} + \frac{S_Y^2}{m}}}.$$
- $T_{n,m}$ will follow asymptotically a standard normal distribution (think about why) so we can compare $T_{n,m}$ to the standard normal to obtain a p-value.

Two-Sample Test: Mean Test - 3

- Sometimes people will compare $T_{n,m}$ to a t-distribution rather than the standard normal distribution. This is because when both samples are from normal distributions (with different means), $T_{n,m}$ has an exact distribution that is the t-distribution.
- When using t-distribution, such a test is called a T-test.
- When using a standard normal distribution, this test is called a Z-test.
- In addition to testing the mean, one can also test other parameters such as median and variance.

Nonparametric Method: KS-test – 1

- Here we introduce a famous test that directly test $H_0 : P_X = P_Y$ – the KS-test.
- The KS-test (Kolmogorov-Smirnov test) is a classical approach in nonparametric two-sample test.

Nonparametric Method: KS-test – 1

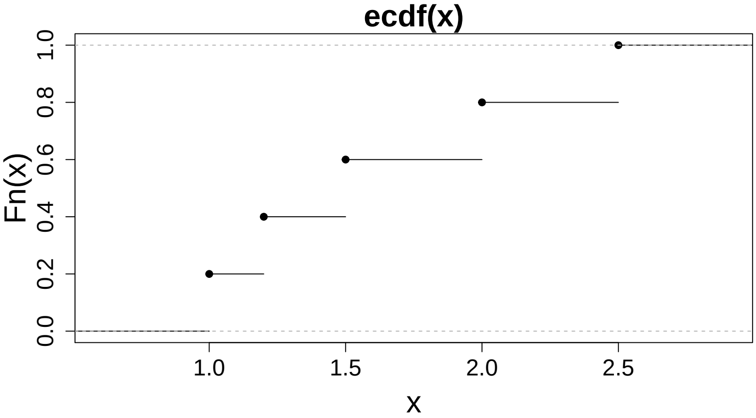
- Here we introduce a famous test that directly test $H_0 : P_X = P_Y$ – the KS-test.
- The KS-test (Kolmogorov-Smirnov test) is a classical approach in nonparametric two-sample test.
- We can estimate the distribution function $P_X(x)$ by the *empirical distribution function (EDF)*:

$$\widehat{P}_X(t) = \frac{1}{n} \sum_{i=1}^{n} I(X_i \leq t),$$

  where $I(x)$ is the indicator function such that if the input is true, then it outputs 1 otherwise 0.
- Actually, $\widehat{P}_X(t)$ is the ratio of data points $X_1, \cdots, X_n$ whose value is less than or equal to t.

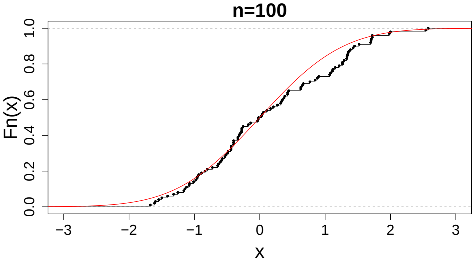
Nonparametric Method: KS-test – 2

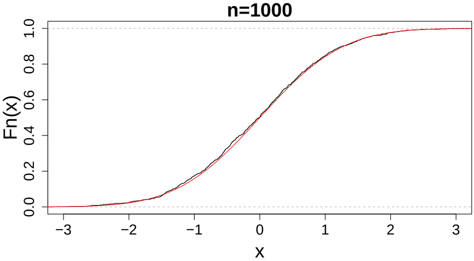
Here is an example of the EDF of 5 observations of $1, 1.2, 1.5, 2, 2.5$:



Nonparametric Method: KS-test – 3

EDF (black curve) versus the true CDF of a standard normal distribution:





Nonparametric Method: KS-Test – 4

- The KS-test is to use the following test statistics:

$$K_{n,m} = \sup_t |\widehat{P}_X(t) - \widehat{P}_Y(t)|,$$

  where $\sup_t$ is a mathematical generalization of $\max_t$ (you can just view it as taking the maximum).
- After rescaling, the test statistics $K_{n,m}$ has a known limiting distribution called the *Kolmogorov distribution*.
- An appealing feature is that the Kolmogorov distribution does not depend on the true distribution P_X and P_Y.
- We then reject the null hypothesis when $K_{n,m}$ is sufficiently large.

Two-Sample Test: Remarks

- The tests we mentioned previously are just common approaches of two-sample test.
- There are many other approaches – if you are interested in, you can search *permutation test*, *rank test, signed-rank test.*
- You can even use histogram to do a two-sample test.
- Keep in mind: there is no universal optimal test and every test works under different assumptions.

Goodness-of-fit Test: χ^2-test - 1

- When we have a theoretical result and we want to compare our data to the theoretical result, we can use the goodness-of-fit test.
- A simple approach to achieve this is through the χ^2 test[1].

[1] here is a tutorial: **http://maxwell.ucsc.edu/~drip/133/ch4.pdf**

Goodness-of-fit Test: χ^2-test - 1

- When we have a theoretical result and we want to compare our data to the theoretical result, we can use the goodness-of-fit test.
- A simple approach to achieve this is through the χ^2 test[1].
- For instance, after some computations, we may obtain the following table (numbers inside parentheses are the SE's):

Value (Errors)	Case 1	Case 2	Case 3	Case 4	Case 5
Observed	16.5 (0.5)	22.1 (0.3)	27.7 (2.2)	25.5 (0.5)	13.2 (0.4)
Theory	15	23	31	25	10

- Can we make some conclusions about the theory using observed data?

[1] here is a tutorial: **http://maxwell.ucsc.edu/~drip/133/ch4.pdf**

Goodness-of-fit Test: χ^2-test - 2

- In the above example, we have 5 statistics $X_1, \cdots, X_5$ (observed value) and each of them has error $\sigma_1, \cdots, \sigma_5$.
- We use $\mu_1, \cdots, \mu_5$ to denote the theoretical result.
- Goal: we want to see if our data fits to the theoretical calculations.
- The χ^2-test compute the test statistic

$$T = \left(\frac{X_1 - \mu_1}{\sigma_1}\right)^2 + \cdots + \left(\frac{X_5 - \mu_5}{\sigma_5}\right)^2.$$

Goodness-of-fit Test: χ^2-test - 3

- If the theory is correct and the noises are independent and Gaussian, then the test statistic T follows a χ^2_5, a χ^2-distribution with 5 degrees of freedom.
- The above example consider 5 statistics. You can generalize it to other number of statistics being considered.

Goodness-of-fit Test: χ^2-test - 3

- If the theory is correct and the noises are independent and Gaussian, then the test statistic T follows a χ^2_5, a χ^2-distribution with 5 degrees of freedom.
- The above example consider 5 statistics. You can generalize it to other number of statistics being considered.
- When reporting the result, you need to state the test statistic (also known as χ^2 statistic) as well as the degree of freedom (number of statistics being compared).
- After computing $T = t$ from the data, the corresponding P-value is $P(\chi^2_\nu \geq t)$ for ν degrees of freedom.

Goodness-of-fit Test: χ^2-test - 3

- If the theory is correct and the noises are independent and Gaussian, then the test statistic T follows a χ^2_5, a χ^2-distribution with 5 degrees of freedom.
- The above example consider 5 statistics. You can generalize it to other number of statistics being considered.
- When reporting the result, you need to state the test statistic (also known as χ^2 statistic) as well as the degree of freedom (number of statistics being compared).
- After computing $T = t$ from the data, the corresponding P-value is $P(\chi^2_\nu \geq t)$ for ν degrees of freedom.
- Note that the null hypothesis being test is

$$H_0 : X_i \sim N(\mu_i, \sigma_i^2), i = 1, \cdots, \nu.$$

Goodness-of-fit Test: χ^2-test - 4

- Recall the observed table:

Value (Errors)	Case 1	Case 2	Case 3	Case 4	Case 5
Observed	16.5 (0.5)	22.1 (0.3)	27.7 (2.2)	25.5 (0.5)	13.2 (0.4)
Theory	15	23	31	25	10

- Thus, in this case, the χ^2 statistic is

$$\begin{aligned}
T &= \left(\frac{X_1 - \mu_1}{\sigma_1}\right)^2 + \cdots + \left(\frac{X_5 - \mu_5}{\sigma_5}\right)^2 \\
&= \left(\frac{16.5 - 15}{0.5}\right)^2 + \left(\frac{22.1 - 23}{0.3}\right)^2 \\
&+ \left(\frac{27.7 - 31}{2.2}\right)^2 + \left(\frac{25.5 - 25}{0.5}\right)^2 + \left(\frac{13.2 - 10}{0.4}\right)^2 \\
&= 3^2 + 3^2 + 1.5^2 + 1^2 + 8^2 \\
&= 85.25
\end{aligned}$$

- Thus, we should report that we observed a signal of a 85.25 χ^2 statistic with 5 degrees of freedom.

Goodness-of-fit Test: Remarks

- When the χ^2 statistic is large, it means that the data *contradicts to the theory or the assumptions about the noises*.
- When the χ^2 statistic is small, it means that the data seems to fit to the theory BUT this does NOT imply that the theory is correct!

Goodness-of-fit Test: Remarks

- When the χ^2 statistic is large, it means that the data *contradicts to the theory or the assumptions about the noises*.
- When the χ^2 statistic is small, it means that the data seems to fit to the theory BUT this does NOT imply that the theory is correct!
- Because the null hypothesis is the theory being true, the goodness-of-fit test can be used to 'prove' that the theoretical result is wrong but it CANNOT be used to prove that the theory is correct!
- Just think about the proof by contradiction: if we cannot show a statement contradicts to itself, this does not imply that statement to be true.

Correlation Test and Independence Test

- There are methods to test if two random variables are correlated. Namely, testing

$$H_0 : \mathbf{corr}(X, Y) = 0,$$

where $\mathbf{corr}(X, Y)$ is the correlation between random variable X and Y.

- Also, we can even test if two random variables are *independent*. Namely, testing

$$H_0 : X \text{ and } Y \text{ are independent. } (\Leftrightarrow p(x, y) = p(x)p(y)).$$

One can use Energy statistics[2] or the approach from reproducing kernel Hilbert space[3].

[2] https://en.wikipedia.org/wiki/Energy_distance#Energy_statistics

[3] http://www.kyb.mpg.de/fileadmin/user_upload/files/publications/attachments/NIPS2007-Gretton_%5b0%5d.pdf

Testing Multiple Hypothesis

- Assume that we are doing hypothesis test 100 times with type-1 errors all being 5%, then it is very likely we are going to reject some H_0's even when they are correct!

---

[4] https://en.wikipedia.org/wiki/Bonferroni_correction

[5] https://en.wikipedia.org/wiki/False_discovery_rate

Testing Multiple Hypothesis

- Assume that we are doing hypothesis test 100 times with type-1 errors all being 5%, then it is very likely we are going to reject some H_0's even when they are correct!
- To avoid falsely rejecting some H_0, we often try to control the *Familywise Error Rate (FWER)*: the chance of falsely rejecting **any** H_0.

[4] https://en.wikipedia.org/wiki/Bonferroni_correction

[5] https://en.wikipedia.org/wiki/False_discovery_rate

Testing Multiple Hypothesis

- Assume that we are doing hypothesis test 100 times with type-1 errors all being 5%, then it is very likely we are going to reject some H_0's even when they are correct!
- To avoid falsely rejecting some H_0, we often try to control the *Familywise Error Rate (FWER)*: the chance of falsely rejecting **any** H_0.
- To make sure FWER is less than α, a classical approach is to use the Bonferroni correction[4]: we only reject those H_0 if their individual p-value is less than α/K where K is the total number of null hypothesis being tested.

---

[4] https://en.wikipedia.org/wiki/Bonferroni_correction

[5] https://en.wikipedia.org/wiki/False_discovery_rate

Testing Multiple Hypothesis

- Assume that we are doing hypothesis test 100 times with type-1 errors all being 5%, then it is very likely we are going to reject some H_0's even when they are correct!
- To avoid falsely rejecting some H_0, we often try to control the *Familywise Error Rate (FWER)*: the chance of falsely rejecting **any** H_0.
- To make sure FWER is less than α, a classical approach is to use the Bonferroni correction[4]: we only reject those H_0 if their individual p-value is less than α/K where K is the total number of null hypothesis being tested.
- There is another criterion that many people commonly use: instead of controlling the FWER, we control the *False Discovery Rate (FDR)*[5].

---

[4] https://en.wikipedia.org/wiki/Bonferroni_correction

[5] https://en.wikipedia.org/wiki/False_discovery_rate

Useful References

- **All of statistics: a concise course in statistical inference.** Larry Wasserman. Springer Science & Business Media, 2013.
- **Statistical inference.** George Casella and Roger L. Berger. Pacific Grove, CA: Duxbury, 2002.
- **Mathematical statistics and data analysis.** John Rice. Nelson Education, 2006.