STAT 231, Final Definitions and Theorems

Descriptive Statistics

Graphical Summaries

1. Histogram

   Standard frequency histogram: height of the rectangle is the frequency or relative frequency.

   Relative frequency histogram: height of the rectangle is relative frequency divided by the length of interval.

2. Empirical cumulative distribution function

3. Boxplots

   $q(0.75) + 1.5 \times IQR, q(0.75), q(0.5), q(0.25), q(0.25) - 1.5 \times IQR$. (The upper(lower) line is placed at the largest observed data value that is smaller than the value $q(0.75) + 1.5 \times IQR$).

4. Qqplots

   Qqplots for checking Gaussian model. All points lie along a string line. Since the quantiles of the Gaussian distribution change in value more rapidly in the tails of the distribution, we expect the points at both ends of the line to lie further from the line.

5. Scatterplots

Numerical Summaries

1. Measure of location.

   Sample mean

   $$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

   Sample median

   Sample mode: highest frequency.

2. Measure of dispersion or variability.

   Sample variance

   $$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} y_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} y_i\right)^2\right]$$

   Range: max - min

   Interquantile range $IQR = q(0.75) - q(0.25)$

3. Measure of Shape

   Sample skewness

   $$g_1 = \frac{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^3}{\left[\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2\right]^{3/2}}$$

   measures the lack of symmetry in the data. Long right tail: positive skewness, long left tail: negative skewness

   Sample kurtosis

   $$g_1 = \frac{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^4}{\left[\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2\right]^{2}}$$

   measures the heaviness of the tails and the peakedness of the data relative to data that are Normally distributed.

4. Sample quantiles and percentiles.

   The 100pth sample percentile is determined as follows

   - $m = (n+1)p$.
   - If m is an integer and $1 \leq m \leq n$, then $q(p) = y(m)$.
   - If m is not an integer but $1 < m < n$, then determine the closest integer j such that $j < m < j+1$ and $q(p) = \frac{1}{2}[y_j + y_{j+1}]$.

5. Sample correlation

$$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}$$

   where

$$S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2$$
$$S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} s_i y_i - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)$$
$$S_{yy} = \sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n} y_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} y_i\right)^2$$

6. Two-way table and Relative risk

   For data

	A	$\bar{A}$	Total
B	y_{11}	y_{12}	$y_{11}+y_{12}$
$\bar{B}$	y_{21}	y_{22}	$y_{21}+y_{22}$
Total	$y_{11}+y_{21}$	$y_{12}+y_{22}$	n

   the relative risk of event A in group B as compared to group $\bar{B}$ is

$$\text{relative risk } = \frac{y_{11}/(y_{11}+y_{12})}{y_{21}/(y_{21}+y_{22})}$$

Statistical Inference

Point Estimation

Definition (Point estimates)**.** A point estimate of a parameter is the value of a function of the observed data $y_1, \ldots, y_n$ and other known quantities such as the sample size n.

Definition (Likelihood function)**.** The likelihood function for θ is defined as

$$L(\theta) = L(\theta; y) = P(Y = y; \theta).$$

Definition (Maximum likelihood estimate)**.** The value of θ which maximizes $L(\theta)$ for given data y is called the maximum likelihood estimate for θ. It is the value of θ which maximizes the probability of observing the data y. This value is denoted $\hat{\theta}$.

Definition (Relative likelihood function)**.** The relative likelihood function is defined as

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

Definition (Log likelihood function)**.** The log likelihood function is defined as

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

Definition (Point estimator and sampling distribution)**.** A point estimator $\tilde{\theta}$ is a random variable which is a function

$$\tilde{\theta} = g(Y_1, \ldots, Y_n).$$

The distribution of $\tilde{\theta}$ is called the sampling distribution of the estimator.

Interval Estimation

Definition (Likelihood interval)**.** A $100p\%$ likelihood interval for θ is the set θ: $R(\theta) \geq p$.

Definition (Confidence interval)**.** A $100p\%$ confidence interval for a parameter is an interval estimate $[L(y), U(y)]$, for which

$$P[L(Y) \leq \theta \leq U(Y)] = p$$

where p is called the confidence coefficient.

The parameter θ is an unknown constant associated with the population. It is not a random variable and therefore does not have a distribution. $L(y), U(y)$ are numerical values not random variables. Hence $P\{\theta \in [L(y), u(y)]\}$ makes no sense.

IMPORTANT: Suppose the experiment which was used to estimate a parameter was conducted a large number of times and each time a 95% confidence interval for the parameter was constructed, then approximately 95% of these constructed intervals would contain the true, but unknown value of the parameter.

Definition (Pivotal Quantities)**.** A pivotal quantity $Q = Q(Y; \theta)$ is a function of the data Y and the unknown parameter θ such that the distribution of the random variable Q is fully known. That is, probability statements such that $P(Q \geq a)$ and $P(Q \leq b)$ depend on a and b but not on θ or any other unknown information.

Theorem. *We can use pivotal quantity to construct confidence interval.*

Proof. Let $P[a \leq Q(Y; \theta) \leq b] = p$ where $Q(Y; \theta)$ is a pivotal quantity whose distribution is completely known. Suppose that we can re-express the inequality $a \leq Q(Y; \theta) \leq b$ in the form $L(Y) \leq \theta \leq U(Y)$ for some functions L and U. Then since

$$\begin{aligned} p &= P[a \leq Q(Y; \theta) \leq b] = P[L(Y) \leq \theta \leq U(Y)] \\ &= P(\theta \in [L(Y), U(Y)]), \end{aligned}$$

the interval $[L(y), U(y)]$ is a $100p\%$ confidence interval for θ. The confidence coefficient θ does not depend on θ. The confidence coefficient θ depends on a and b, and these are determined by the known distribution of $Q(Y; \theta)$. □

Example (Confidence interval for the mean μ of a Gaussian distribution with known standard deviation σ)**.** Suppose $Y = (Y_1, \ldots, Y_n)$ is a random sample from the $G(\mu, \sigma)$ distribution where $E(Y_i) = \mu$ is unknown but $sd(Y_i) = \sigma$ is known. Since

$$Q = Q(Y; \mu) = \frac{\bar{Y} - \mu}{\sigma/\sqrt{n}} \sim G(0, 1)$$

and $G(0, 1)$ is a completely known distribution, Q is a pivotal quantity.

Example (Approximate confidence interval for Binomial model)**.** For large n, denote $Y = \sum_{i=1}^{n} Y_i$,

$$Q_n = Q_n(Y; \theta) = \frac{Y - n\theta}{[n\tilde{\theta}(1 - \tilde{\theta})]^{1/2}}$$

where $\tilde{\theta} = Y/n$, is also close to $G(0, 1)$. Thus

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

gives an approximate $100p\%$ confidence interval for θ where $p = P(-a \leq Z \leq a), Z \sim G(0, 1)$.

Definition (The χ^2 Distribution)**.** The $\chi^2(k)$ distribution is a continuous family of distributions on $(0, \infty)$ with probability density function of the form

$$f(x; k) = \frac{1}{2^{k/2}\Gamma(k/2)} x^{(k/2)-1} e^{-x/2}$$

where $k \in \{1, 2, \ldots\}$ is the degrees of freedom parameter.

For $k = 2$, the probability density function is the $Exponential(2)$ probability density function.

For $k > 2$, the probability density function has maximum value at $x = k - 2$.

For $k \geq 30$, the probability density function resembles that of a $N(k, 2k)$ probability density function.

Theorem. *If $Z \sim G(0, 1)$ then the distribution $W = Z^2$ is $\chi^2(1)$.*

Corollary. *If $W \sim \chi^2(1)$ then $P(W \geq w) = 2[1 - P(Z \leq \sqrt{w})]$ where $Z \sim G(0,1)$.*

Definition (Student's t distribution)**.** Student's t distribution has probability density function

$$f(t;k) = c_k(1 + \frac{t^2}{k})^{-(k+1)/2}$$

where the constant c_k is given by

$$c_k = \frac{\Gamma(\frac{k+1}{2})}{\sqrt{k\pi}\Gamma(\frac{k}{2})}$$

The parameter k is called the degrees of freedom. The t probability density function is symmetric about the origin, and for large values of k, the graph of the probability density function $f(t;k)$ is indistinguishable from that of the $G(0,1)$ probability density function.

Theorem. *Suppose $Z \sim G(0,1)$ and $U \sim \chi^2(k)$ independently. Let*

$$T = \frac{Z}{\sqrt{U/k}}$$

Then T has a Student's t distribution with k degrees of freedom.

Theorem. *If $L(\theta)$ is based on $Y = (Y_1, \ldots, Y_n)$, a random sample of size n, and if θ is the true value of the scalar parameter, then the distribution of $\Lambda(\theta)$ converges to a $\chi^2(1)$ distribution as $n \to \infty$ where*

$$\Lambda(\theta) = -2\log\left[\frac{L(\theta)}{L(\hat{\theta})}\right]$$

Theorem. *A $100p\%$ likelihood interval is an approximate $100q\%$ confidence interval where $q = 2P(Z \leq \sqrt{-2\log p}) - 1$ and $Z \sim N(0,1)$.*

Proof. First,

$$\{\theta; R(\theta) \geq p\} = \left\{\theta\colon\ -2\log\left[\frac{L(\theta)}{L(\hat{\theta})}\right] \leq -2\log p\right\}$$

Then

$$\begin{aligned} P[\Lambda(\theta) \leq -2\log p] &= P\{-2\log\left[\frac{L(\theta)}{L(\hat{\theta})}\right] \leq -2\log p\} \\ &\approx P(W \leq -2\log p) \text{ where } W \sim \chi^2(1) \\ &= 2P(Z \leq \sqrt{-2\log p}) - 1 \text{ where } Z \sim N(0,1) \end{aligned}$$

as required. □

Tests of Hypotheses

Definition (Null and alternative hypotheses)**.** The default hypothesis is often referred to as the null hypothesis and is denoted by H_0. The alternative hypothesis is, in many cases, that H_0 is not true.

Definition (Test statistic)**.** A test statistic D is a function of the data Y that is constructed to measure the degree of agreement between the data Y and the null hypothesis H_0.

Definition (p-value)**.** Suppose we use the test statistic $D = D(Y)$ to test the hypothesis H_0. Suppose also that $d = D(y)$ is the observed value of D. The p-value or observed significance level of the test hypothesis H_0 using test statistic D is

$$p - value = P(D \geq d; H_0)$$

Theorem (Relationship between hypothesis testing and interval estimation)**.** *Suppose we have data y, a model $f(y;\theta)$ and we use the same pivotal quantity to construct a confidence interval for θ and a test for the hypothesis $H_0\colon \theta = \theta_0$. Then the parameter value $\theta = \theta_0$ is inside a $100p\%$ confidence interval for θ if and only if the p-value for testing $H_0\colon \theta = \theta_0$ is greater than $1 - q$.*

Statistical Models

Binomial

1. Likelihood function

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

$$\hat{\theta} = y/n$$

2. Approximate Confidence interval

   Pivotal Quantity

$$Q_n = Q_n(Y;\theta) = \frac{Y - n\theta}{\left[n\tilde{\theta}(1-\tilde{\theta})\right]^{1/2}} \sim G(0,1)$$

   Interval

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

3. Likelihood Ratio Test of Hypothesis - One Parameter

   Hypothesis

$$H_0 \colon \theta = \theta_0$$

   Likelihood ratio statistic

$$\Lambda(\theta_0) = -2\log\left[\frac{L(\theta_0)}{L(\tilde{\theta})}\right] \sim \chi^2(1)$$

   p-value

$$\text{p-value} = P[W \geq \lambda] = 2[1 - P(Z \leq \sqrt{\lambda})]$$

Poisson

1. Likelihood function

$$L(\theta) = \theta^{n\bar{y}} e^{-n\theta}$$

$$\hat{\theta} = \bar{y}$$

Geometric

1. Likelihood function

$$L(\theta) = \frac{1}{\theta^n}\left(-\sum_{i=1}^{n} y_i/\theta\right)$$

$$\hat{\theta} = \bar{y}$$

Multinomial

1. Likelihood function

$$L(\theta) = \prod_{i=1}^{k} \theta_i^{y_i}$$

$$\hat{\theta}_i = \frac{y_i}{n}$$

Gaussian

1. Likelihood function

$$L(\theta) = L(\mu, \sigma) = \sigma^{-n} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mu)^2\right]$$

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} y_i = \bar{y}$$

$$\hat{\sigma} = [\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2]^{1/2}$$

2. Confidence interval

   (a) Known standard deviation
   Pivotal Quantity

$$Q = Q(Y; \mu) = \frac{\bar{Y} - \mu}{\sigma/\sqrt{n}} \sim G(0, 1)$$

   Interval

$$[\bar{y} - a\sigma/\sqrt{n}, \bar{y} + a\sigma/\sqrt{n}]$$

   (b) Unknown σ
   Pivotal quantity for μ

$$T = \frac{\bar{Y} - \mu}{S/\sqrt{n}} \sim t(n-1)$$

   Confidence interval for μ

$$\bar{y} \pm as/\sqrt{n}$$

   Pivotal quantity for σ^2

$$\frac{(n-1)S^2}{\sigma^2} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(Y_i - \bar{Y})^2 \sim \chi^2(n-1)$$

   Confidence interval for σ^2

$$\left[\sqrt{\frac{(n-1)s^2}{b}}, \sqrt{\frac{(n-1)s^2}{a}}\right]$$

   NOTE: The choice of a, b is not unique. For convenience, a and b are usually chosen such that

$$P(U \leq a) = P(U \geq b) = \frac{1-p}{2}$$

3. Prediction Interval for a Future Observation
   Pivotal quantity for future observation

$$\frac{Y - \bar{Y}}{S\sqrt{1 + \frac{1}{n}}} \sim t(n-1)$$

   Interval for future observation

$$\left[\bar{y} - as\sqrt{1 + \frac{1}{n}}, \bar{y} + as\sqrt{1 + \frac{1}{n}}\right]$$

4. Hypotheses testing

   (a) Test of Hypothesis for μ
   Hypothesis

$$H_0 \colon \mu = \mu_0$$

   Test statistic

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

   p-value

$$\text{p-value} = P(D \geq d; H_0 = true) = P(|T| \geq d) = 2[1 - P(T \leq d)]$$

   This is called two-sided test. There is also one-sided test. Simply remove absolute sign in D, and $p - value = 1 - P(T \leq d)$.

(b) Test of Hypothesis for σ

Hypothesis

$$H_0 : \sigma = \sigma_0$$

Test statistic

$$U = \frac{(n-1)S^2}{\sigma_0^2} \sim \chi^2(n-1)$$

p-value

- If u is large, that is $P(U \leq u) > 1/2$,

$$\text{p-value} = 2P(U \geq u)$$

- if u is small, that is $P(U \leq u) < 1/2$,

$$\text{p-value} = 2P(U \leq u)$$

Gaussian response model

Simple linear regression

$$Y_i = G(\mu(x_i), \sigma) \text{ where } \mu(x_i) = \alpha + \beta x_i$$

1. Estimator

   - Maximum likelihood estimators

$$\tilde{\beta} = \frac{S_{xy}}{S_{xx}}$$
$$\tilde{\alpha} = \bar{Y} - \tilde{\beta}\bar{x}$$
$$\tilde{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \tilde{\alpha} - \tilde{\beta}x_i)^2$$

   - Least square estimator, same as maximum likelihood estimators.

2. Confidence interval

   - Distribution of the estimator $\tilde{\beta}$

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

   Pivotal quantity to obtain confidence intervals for β

$$\frac{\tilde{\beta} - \beta}{S_e/\sqrt{S_{xx}}} \sim t(n-2)$$

   Interval

$$\hat{\beta} \pm a s_e/\sqrt{S_{xx}}$$

   Hypothesis of no relationship

$$H_0 : \beta = 0$$

   Test statistic

$$\frac{\left|\tilde{\beta} - 0\right|}{S_e/\sqrt{S_{xx}}}$$

   p-value, $T \sim t(n-2)$

$$P\left(|T| \geq \frac{\left|\hat{\beta} - 0\right|}{s_e/\sqrt{S_{xx}}}\right)$$

- Confidence interval for the mean response $\mu(x) = \alpha + \beta x$
  Distribution
  $$\tilde{\mu}(x) \sim G\left(\mu(x), \sigma\sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}}\right)$$
  Pivotal quantity
  $$\frac{\tilde{\mu}(x) - \mu(x)}{S_e\sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}}} \sim t(n-2)$$
  Interval
  $$\hat{\mu}(x) \pm as_e\sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}}$$
- Prediction Interval for future response
  Distribution
  $$\tilde{\mu}(x) \sim G\left(\mu(x), \sigma\sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}}\right)$$
  Pivotal quantity
  $$\frac{\tilde{\mu}(x) - \mu(x)}{S_e\sqrt{1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}}} \sim t(n-2)$$
  Interval
  $$\hat{\mu}(x) \pm as_e\sqrt{1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}}$$

3. Comparing mean of two population

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

$$\frac{(\bar{Y_1} - \bar{Y_2}) - (\mu_1 - \mu_2)}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t(n_1+n_2-2)$$

$$\frac{(n_1+n_2-2)S_p^2}{\sigma^2} = \frac{1}{\sigma^2}\sum_{j=1}^{2}\sum_{i=1}^{n_j}(Y_{ji} - \bar{Y_j})^2 \sim \chi^2(n_1+n_2-2)$$

$$\bar{y_1} - \bar{y_2} \pm as_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

Goodness of git

1. Multinomial model

$$\Lambda = 2\sum_{j=1}^{k} Y_j \log\left(\frac{Y_j}{E_j}\right)$$

$$\text{p-value} = P(\Lambda \geq \lambda; H_0) \approx P(W \geq \lambda), W \sim \chi^2(k-1-p)$$

Pearson goodness of git statistic

$$D = \sum_{j=1}^{k}\frac{(Y_j - E_j)^2}{E_j}$$

2. two way table

   Likelihood ratio statistic

$$2\sum_{i=1}^{m}\sum_{j=1}^{n} Y_{ij}\log\left(\frac{Y_{ij}}{E_{ij}}\right)$$

   Distribution is $\chi^2((a-1)(b-1))$