Populations and Samples
BioS 662

Michael G. Hudgens, Ph.D.
mhudgens@bios.unc.edu

http://www.bios.unc.edu/~mhudgens

2008-08-22 16:29
Random Variables

- **Random sample**: result of independently selecting elements at random from a population

- Def 4.8 A *random variable* is a variable associated with a random sample

PV Rao (p 786, 1998) A *rv* is a variable whose value is determined by the observed characteristics of an item randomly selected from a population
Probability Functions

- Def 4.9 The *probability mass function* is a function that for each possible value of a discrete rv takes on the probability of that value occurring.

- Def 4.10 The *probability density function* is a curve that specifies, by means of the area under the curve over an interval, the probability that a continuous rv falls within the interval.
Probability Functions

Number of Boys in Families with Eight Children

Probability mass function

Probability density function
Cumulative Distribution Function

- Def 4.9 The *cumulative distribution function* for a rv $X$ is

  $$ F(x) = \Pr[X \leq x] $$

- If $X$ is discrete,

  $$ F(x) = \sum_{y \leq x} p_X(y) $$

  where $p_X$ is the pmf of $X$

- If $X$ is continuous,

  $$ F(x) = \int_{-\infty}^{x} f(y) \, dy $$

  where $f$ is the pdf of $X$
Population quantile

• Intuitive definition:

  The $p^{th}$ quantile of $X$, say $\zeta_p$, should be such that

  $$F(\zeta_p) = \Pr[X \leq \zeta_p] = p$$

• Formally:

  $$\zeta_p = \inf\{x : F(x) \geq p\}$$

• If $F$ is continuous

  $$F(\zeta_p) = p$$
Quantiles: Example

![Graphical representation of quantiles]

BIOS 662

Populations and Samples
Mean and Variance

• Mean or expected value of $X$
  
  - If $X$ is discrete,
    \[
    \mu = E(X) = \sum_{y} yp_X(y)
    \]
  
  - If $X$ is continuous,
    \[
    \mu = E(X) = \int_{-\infty}^{\infty} yf(y)dy
    \]

• Variance
  \[
  \sigma^2 = Var(X) = E\{(X - \mu)^2\} \]
Skewness and Kurtosis

- **Skewness**
  \[ \alpha_3 = \frac{E\{(X - \mu)^3\}}{\sigma^3} \]

- **Kurtosis**
  \[ \alpha_4 = \frac{E\{(X - \mu)^4\}}{\sigma^4} \]
Parameters and Statistics

• Definition: A *parameter* is a numerical characteristic of a population

• Definition: A *statistic* is a numerical characteristic of a sample

• Notation: Greek letters typically denote parameters; English letters denote statistics

• Example:

\[ \mu = \text{population mean}; \ \sigma^2 \text{ population variance} \]

\[ \bar{Y} = \text{sample mean}; \ s^2 \text{ sample variance} \]
Parameters and Statistics

- Parameters are fixed constants
- Statistics are random variables
- Statistics have probability distributions
- We will use statistics and probability theory to draw conclusions (inference) about parameters
Sampling Distributions

• Definition 4.15 The probability function of a statistic is called the *sampling distribution of the statistic*.

• Eg, when sampling from a population, the sample mean \( \bar{Y} \) is a rv becuase its value depends on chance, namely, on which sample is obtained.

The probability distribution of the random variable \( \bar{Y} \) is called the *sampling distribution of the mean*. 
Sampling Distributions

• Result 4.1 If a rv $Y$ has a population mean $\mu$ and a population variance $\sigma^2$, the sampling distribution of the mean ($\bar{Y}$) has mean $\mu$ and variance $\sigma^2/n$

• Definition 4.16 The standard deviation of the sampling distribution is called the \textit{standard error}

• Eg the standard error of $\bar{Y}$ is $\sigma/\sqrt{n}$
Normal or Gaussian Distribution

- **PDF:**
  
  \[ f(x; \mu, \sigma) = \frac{1}{\sigma \sqrt{2\pi}} \exp \left\{ -\frac{1}{2} \left( \frac{x - \mu}{\sigma} \right)^2 \right\} \]

- **CDF:**
  
  \[ F(x; \mu, \sigma) = \int_{-\infty}^{x} f(y; \mu, \sigma) \, dy \]

- **\( \mu \)** mean, **\( \sigma^2 \)** variance
- **\( X \sim N(\mu, \sigma^2) \)** [beware \( X \sim N(\mu, \sigma) \)]
Normal Distribution

The image shows several normal distribution curves labeled as:
- \( N(25, 900) \)
- \( N(25, 225) \)
- \( N(100, 81) \)
- \( N(100, 225) \)

The x-axis represents the variable \( x \) ranging from -50 to 150, and the y-axis represents the density ranging from 0.00 to 0.05.

The curves illustrate how different parameters affect the shape and spread of normal distributions.
Standard Normal Distribution

- $Z \sim N(0, 1)$
- PDF:
  \[
  \phi(z) = \frac{1}{\sqrt{2\pi}} \exp\left\{ -\frac{1}{2}z^2 \right\}
  \]
- CDF:
  \[
  \Phi(z) = \int_{-\infty}^{z} \phi(y)\,dy
  \]
- $N(0, 1)$ is a standard normal distribution
Standard Normal Distribution

\begin{figure}
\centering
\includegraphics[width=\textwidth]{standard_normal_pdf}
\caption{Standard Normal Distribution}
\end{figure}
Properties of Standard Normal Distribution

- A rv w/ pdf $f$ is symmetric about $\mu$ if
  \[ f(\mu + x) = f(\mu - x) \text{ for all } x \]

- $Z \sim N(0, 1)$ is symmetric about 0
  \[ \phi(z) = \phi(-z) \text{ for all } -\infty < z < \infty \]

- Thus
  \[ \Pr[Z \leq -z] = \Pr[Z \geq z] \]

  i.e.
  \[ \Phi(-z) = 1 - \Phi(z) \]
Standard Normal Distribution

\[ Pr[Z < -0.1] = 0.4602 \]
\[ Pr[0.1 < Z] = 0.4602 \]
Standard Normal Distribution

- R

```r
> pnorm(-.1,0,1)
[1] 0.4601722
> 1 - pnorm(.1,0,1)
[1] 0.4601722
> qnorm(0.4601722,0,1)
[1] -0.0999999
```
Standard Normal Distribution

- SAS

```sas
data;
    x=probnorm(-.1);
    y=cdf('NORMAL',-.1,0,1);
    z=quantile('NORMAL',0.4601722);
run;

proc print; run;
```

<table>
<thead>
<tr>
<th>Obs</th>
<th>x</th>
<th>y</th>
<th>z</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.46017</td>
<td>0.46017</td>
<td>-0.100000</td>
</tr>
</tbody>
</table>
Properties of a Random Variable

- Let $X$ be a random variable
- Suppose $Y = aX + b$ where $a$ and $b$ are constants
- Then
  \[ E(Y) = aE(X) + b \]
  \[ Var(Y) = a^2 Var(X) \]
- If $X \sim N(\mu, \sigma^2)$ and $Y = aX + b$, then
  \[ Y \sim N(a\mu + b, (a\sigma)^2) \]
Conversion to Standard Normal

• Suppose \( Y \sim N(\mu, \sigma^2) \)

• Let

\[
Z = \frac{Y - \mu}{\sigma}
\]

• Then

\( Z \sim N(0, 1) \)

• In words: a normal random variable can be standardized by subtracting its mean and dividing by its standard deviation
Computation of Probabilities

• Suppose $Y \sim N(\mu, \sigma^2)$

• Let

$$Z = \frac{Y - \mu}{\sigma}$$

• Then

$$Pr[a < Y < b] = Pr\left[\frac{a - \mu}{\sigma} < Z < \frac{b - \mu}{\sigma}\right]$$

$$= \Phi\left(\frac{b - \mu}{\sigma}\right) - \Phi\left(\frac{a - \mu}{\sigma}\right)$$
Table 1 (p 818 text): Standard normal distribution.

Let $Z$ be a normal random variable with mean zero and variance one. For selected values of $z$, three values are tabled: (1) the two-sided $p$-value, or $\Pr(|Z| \geq z)$; (2) the one-sided $p$-value, or $\Pr[Z \geq z]$; and (3) the cumulative distribution function at $z$, or $\Pr[Z \leq z]$.

<table>
<thead>
<tr>
<th>$z$</th>
<th>Two sided</th>
<th>One sided</th>
<th>Cumu. dist.</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.00</td>
<td>1.0000</td>
<td>.5000</td>
<td>.5000</td>
</tr>
<tr>
<td>0.05</td>
<td>.9601</td>
<td>.4801</td>
<td>.5199</td>
</tr>
<tr>
<td>0.10</td>
<td>.9203</td>
<td>.4602</td>
<td>.5398</td>
</tr>
<tr>
<td>0.15</td>
<td>.8808</td>
<td>.4404</td>
<td>.5596</td>
</tr>
<tr>
<td>0.20</td>
<td>.8415</td>
<td>.4207</td>
<td>.5793</td>
</tr>
<tr>
<td>0.25</td>
<td>.8026</td>
<td>.4013</td>
<td>.5987</td>
</tr>
<tr>
<td>0.30</td>
<td>.7642</td>
<td>.3821</td>
<td>.6179</td>
</tr>
<tr>
<td>0.35</td>
<td>.7263</td>
<td>.3632</td>
<td>.6368</td>
</tr>
<tr>
<td>0.40</td>
<td>.6892</td>
<td>.3446</td>
<td>.6554</td>
</tr>
<tr>
<td>0.45</td>
<td>.6527</td>
<td>.3264</td>
<td>.6736</td>
</tr>
<tr>
<td>0.50</td>
<td>.6171</td>
<td>.3085</td>
<td>.6915</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.00</td>
<td>.3173</td>
<td>.1587</td>
<td>.8413</td>
</tr>
<tr>
<td>1.33</td>
<td>.1835</td>
<td>.0918</td>
<td>.9082</td>
</tr>
<tr>
<td>1.64</td>
<td>.1010</td>
<td>.0505</td>
<td>.9495</td>
</tr>
<tr>
<td>1.96</td>
<td>.0500</td>
<td>.0250</td>
<td>.9750</td>
</tr>
<tr>
<td>2.00</td>
<td>.0455</td>
<td>.0288</td>
<td>.9772</td>
</tr>
<tr>
<td>2.58</td>
<td>.0099</td>
<td>.0049</td>
<td>.9951</td>
</tr>
</tbody>
</table>
Example

- Intraocular pressure (IP) is used to diagnose glaucoma
- Assume IP is normally distributed with mean $\mu = 16$ mmHg and variance $\sigma^2 = 9$ mmHg
- If pressure greater than 20 mmHg is considered abnormal, what proportion of the population is abnormal?

$$
Pr[X > 20] = Pr\left[\frac{X-16}{3} > \frac{20-16}{3}\right]
$$

$$
= Pr[Z > 1.33] = 1 - \Phi(1.33)
$$

$$
= 1 - 0.9082 = 0.0918
$$
Example (continued)

- What proportion of the population has IP between 4 and 18?

\[
\begin{align*}
\Pr[4 < X < 18] &= \Pr\left[\frac{4-16}{3} < \frac{X-16}{3} < \frac{18-16}{3}\right] \\
&= \Pr[-4 < Z < 2/3] \\
&= \Phi\left(\frac{2}{3}\right) - \Phi(-4) \\
&= \Phi\left(\frac{2}{3}\right) - 1 + \Phi(4) \\
&= 0.7486 - 1 + 0.9999 = 0.7485
\end{align*}
\]
Assessing Normality

• How do we assess whether the normal distribution model fits a particular set of data?

• One graphical approach: quantile-quantile (QQ) plot

• Plot quantiles of the observed data distribution versus the quantiles of the normal distribution

• Straight line indicates normality assumption reasonable
QQ Plot Example

- Table 4.3 from text (p. 81)

<table>
<thead>
<tr>
<th>Endpoint</th>
<th>Freq</th>
<th>Pct</th>
</tr>
</thead>
<tbody>
<tr>
<td>61.7</td>
<td>5</td>
<td>0.5</td>
</tr>
<tr>
<td>62.2</td>
<td>7</td>
<td>1.3</td>
</tr>
<tr>
<td>63.2</td>
<td>32</td>
<td>4.7</td>
</tr>
<tr>
<td>64.2</td>
<td>59</td>
<td>11.1</td>
</tr>
<tr>
<td>65.2</td>
<td>48</td>
<td>16.3</td>
</tr>
<tr>
<td>66.2</td>
<td>117</td>
<td>28.9</td>
</tr>
<tr>
<td>67.2</td>
<td>138</td>
<td>43.8</td>
</tr>
<tr>
<td>68.2</td>
<td>120</td>
<td>56.7</td>
</tr>
<tr>
<td>69.2</td>
<td>167</td>
<td>74.7</td>
</tr>
<tr>
<td>70.2</td>
<td>99</td>
<td>85.3</td>
</tr>
<tr>
<td>71.2</td>
<td>64</td>
<td>92.2</td>
</tr>
<tr>
<td>72.2</td>
<td>41</td>
<td>96.7</td>
</tr>
<tr>
<td>73.2</td>
<td>17</td>
<td>98.5</td>
</tr>
<tr>
<td>73.7</td>
<td>14</td>
<td>100</td>
</tr>
</tbody>
</table>
QQ Plot Example: Table 4.3 from text

> qqnorm(galton,type="l")
Sample Quantiles

Theoretical Quantiles

QQ Plot

> par(mfcol=c(1,2)); qqnorm(rnorm(1000,0,1)); qqnorm(rexp(1000,3))
> x <- rexp(1000, 3)
> par(mfcol=c(1,2)); qqplot(rexp(1000,3),x); qqplot(rnorm(1000,0,1),x)
> x <- rexp(1000,3); probs <- seq(.01,.99,length=100)
> qx <- quantile(x,probs); tqexp <- qexp(probs,3); tqnorm <- qnorm(probs,0,1)
> par(mfcol=c(1,2)); plot(tqexp,qx,xlab="Theoretical quantiles of Exp(1/3)",ylab="Sample quantiles")
> plot(tqnorm,qx,xlab="Theoretical quantiles of N(0,1)",ylab="Sample quantiles")
Some approximations

• The interval $\bar{x} \pm s$ will contain approx 68% of the observations

• The interval $\bar{x} \pm 2s$ will contain approx 95% of the observations

• Assuming $Y \sim N(\mu, \sigma^2)$

  \[
  \Pr[\mu - \sigma < Y < \mu + \sigma] = \Pr[-1 < Z < 1] = 0.6827
  \]

  \[
  \Pr[\mu - 2\sigma < Y < \mu + 2\sigma] = \Pr[-2 < Z < 2] = 0.9545
  \]
Some approximations

• Do these approximations hold for non-normal data?
  Not in general.

• Consider $X \sim Exp(\lambda)$ such that $EX = \lambda$ and $V(X) = \lambda^2$. For $\lambda = 1/3$,

$$\Pr[0 \leq X \leq 2/3] = 0.86$$
Some approximations

• Consider \( X = WY + (1-W)Z \) where \( W \sim Bern(1/2) \), 
  \( Y \sim N(10, 1) \), and \( Z \sim N(0, 1) \). Can show

\[
\Pr[\mu_X - \sigma_X \leq X \leq \mu_X + \sigma_X] \approx 0.54
\]
Some approximations

- The following holds for any rv $Y$ with mean $\mu$ and variance $\sigma^2$

$$\Pr[\mu - K\sigma \leq Y \leq \mu + K\sigma] \geq 1 - \frac{1}{K^2}$$

$\forall K \geq 1$. This is Chebyshev’s inequality (note typo in text page 100)

- Eg, if $K = 2$,

$$\Pr[\mu - 2\sigma \leq Y \leq \mu + 2\sigma] \geq 0.75$$

i.e. we would expect at least 75% of observations to be within two standard deviations of the mean for any underlying distribution
Central Limit Theorem (CLT)

- Let \( Y_1, Y_2, \ldots, Y_n \) be independent and identically distributed (iid) random variables with
  
  \[
  E(Y_i) = \mu \quad \text{and}, 
  \]
  
  \[
  Var(Y_i) = \sigma^2 > 0 
  \]

- Define

  \[
  Z_n = \frac{\bar{Y} - \mu}{\sigma / \sqrt{n}} 
  \]

- Then the distribution function of \( Z_n \) converges to a standard normal distribution function as \( n \rightarrow \infty \).
Central Limit Theorem (CLT)

- In words - see Result 4.3 page 84 of van Belle et al

- If a random variable $Y$ has population mean $\mu$ and population variance $\sigma^2$, then the sample mean $\bar{Y}$, based on $n$ observations, is approx normally distributed with mean $\mu$ and variance $\sigma^2/n$ for sufficiently large $n$
Notes on CLT

• The CLT applies to any distribution of the $Y$’s
• The approximation improves as $n$ gets large
• Check out Rice Virtual Lab in Statistics

http://onlinestatbook.com/rvls.html
Result 4.2

- If $Y$ is normally distributed with mean $\mu$ and variance $\sigma^2$, then $\bar{Y}$, based on a random sample of $n$ observations, is normally distributed with mean $\mu$ and variance $\sigma^2/n$.
- This is true regardless of sample size.