Chapter 1: Descriptive statistics

August 31st, 2017

Topics

Week 1	**Chapter 1: Descriptive statistics**
Week 2	Chapter 6: Statistics and Sampling Distributions
Week 4	Chapter 7: Point Estimation
Week 7	Chapter 8: Confidence Intervals
Week 10	Chapter 9: Test of Hypothesis
Week 13	Two-sample inference, ANOVA, regression

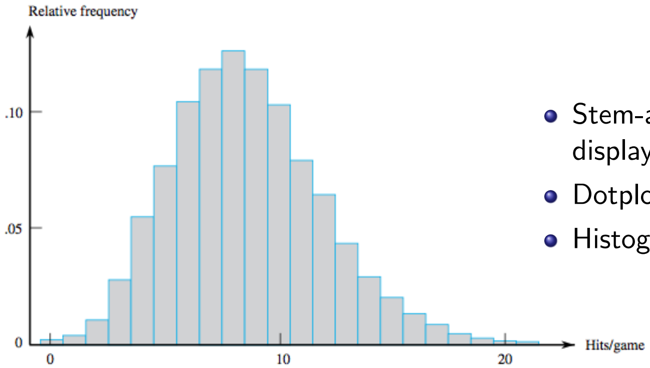


Figure 1.6 Histogram of number of hits per nine-inning game

- Stem-and-Leaf displays
- Dotplots
- Histograms

1.3: Measures of locations

- The Mean
- The Median
- Trimmed Means

Measures of locations: mean

The **sample mean** $\bar{x}$ of observations $x_1, x_2, \ldots, x_n$ is given by

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Measures of locations: median

Step 1: ordering the observations from smallest to largest

$$\tilde{x} = \begin{cases} \text{The single middle value if } n \text{ is odd} & = \left(\dfrac{n+1}{2}\right)^{\text{th}} \text{ordered value} \\ \text{The average of the two middle values if } n \text{ is even} & = \text{average of } \left(\dfrac{n}{2}\right)^{\text{th}} \text{ and } \left(\dfrac{n}{2}+1\right)^{\text{th}} \text{ ordered values} \end{cases}$$

Median is not affected by outliers

Measures of locations: trimmed mean

- A α% trimmed mean is computed by:
  - eliminating the smallest α% and the largest α% of the sample
  - averaging what remains
- $\alpha = 0 \rightarrow$ the mean
- $\alpha \approx 50 \rightarrow$ the median

Measures of Variability: deviations from the mean

The **sample variance**, denoted by s^2, is given by

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n - 1} = \frac{S_{xx}}{n - 1}$$

The **sample standard deviation**, denoted by s, is the (positive) square root of the variance:

$$s = \sqrt{s^2}$$

Measures of Variability: deviations from the mean

The **sample variance**, denoted by s^2, is given by

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n - 1} = \frac{S_{xx}}{n - 1}$$

The **sample standard deviation**, denoted by s, is the (positive) square root of the variance:

$$s = \sqrt{s^2}$$

- Why squared? Because it is easier to do math with x^2 than $|x|$
- Why $(n-1)$? Because that makes s^2 an *unbiased estimator* of the population variance (Chapter 7)

Computing formula for s^2

$$S_{xx} = \sum (x_i - \bar{x})^2 = \sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}$$

Proof Because $\bar{x} = \sum x_i/n$, $n\bar{x}^2 = \left(\sum x_i\right)^2/n$. Then,

$$\sum (x_i - \bar{x})^2 = \sum (x_i^2 - 2\bar{x} \cdot x_i + \bar{x}^2) = \sum x_i^2 - 2\bar{x}\sum x_i + \sum (\bar{x})^2$$

$$= \sum x_i^2 - 2\bar{x} \cdot n\bar{x} + n(\bar{x})^2 = \sum x_i^2 - n(\bar{x})^2$$

Properties of the sample standard deviation

Let $x_1, x_2, \ldots, x_n$ be a sample and c be a constant.

1. If $y_1 = x_1 + c, y_2 = x_2 + c, \ldots, y_n = x_n + c$, then $s_y^2 = s_x^2$, and

2. If $y_1 = cx_1, \ldots, y_n = cx_n$, then $s_y^2 = c^2 s_x^2$, $s_y = |c| s_x$,

where s_x^2 is the sample variance of the x's and s_y^2 is the sample variance of the y's.

Boxplots

Order the n observations from smallest to largest and separate the smallest half from the largest half; the median $\tilde{x}$ is included in both halves if n is odd. Then the **lower fourth** is the median of the smallest half and the **upper fourth** is the median of the largest half. A measure of spread that is resistant to outliers is the **fourth spread** f_s, given by

$$f_s = \text{upper fourth} - \text{lower fourth}$$

Boxplots

$$\underbrace{40\quad 52\quad 55\quad 60\quad 70\quad 75\quad 85\quad 85\quad 90}\quad \overbrace{90\quad 92\quad 94\quad 94\quad 95\quad 98\quad 100\quad 115\quad 125\quad 125}$$

The five-number summary is as follows:

smallest $x_i = 40$ lower fourth $= 72.5$ $\tilde{x} = 90$ upper fourth $= 96.5$
largest $x_i = 125$

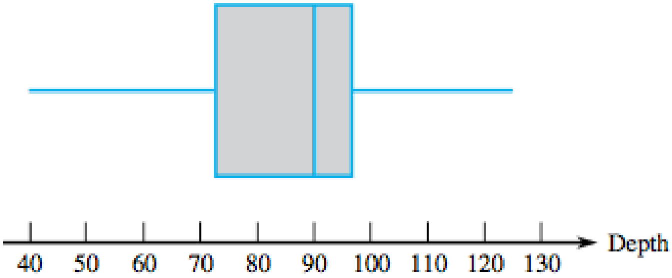
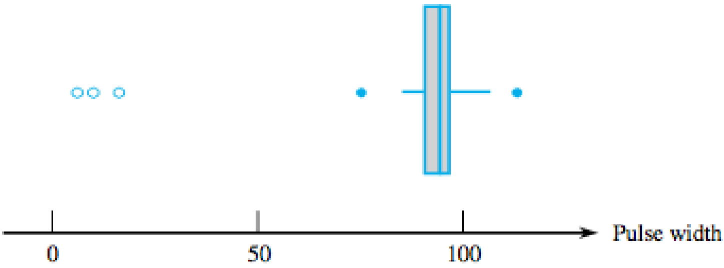
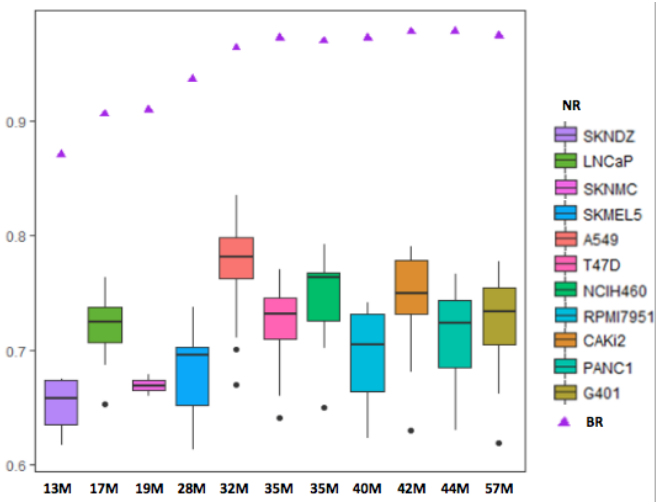


Figure 1.17 A boxplot of the corrosion data

Any observation farther than $1.5f_s$ from the closest fourth is an **outlier**. An outlier is **extreme** if it is more than $3f_s$ from the nearest fourth, and it is **mild** otherwise.



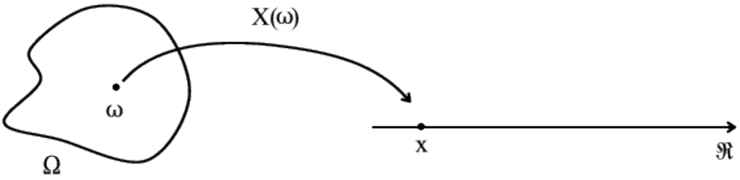
Comparative boxplots



Review: Random variables, expected values, normal distribution

Reading: 3.1, 3.2, 3.3, 4.1, 4.2

Random variable



x	1	2	3	4	5	6	7
$p(x)$	.01	.03	.13	.25	.39	.17	.02

Random variable (continuous)

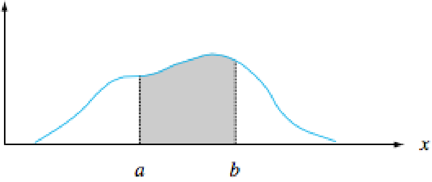


Figure 4.2 $P(a \le X \le b)$ = the area under the density curve between a and b

Let X be a continuous rv. Then a **probability distribution** or **probability density function** (pdf) of X is a function $f(x)$ such that for any two numbers a and b with $a \le b$,

$$P(a \le X \le b) = \int_a^b f(x)\,dx$$

Cumulative distribution function

The **cumulative distribution function** $F(x)$ for a continuous rv X is defined for every number x by

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

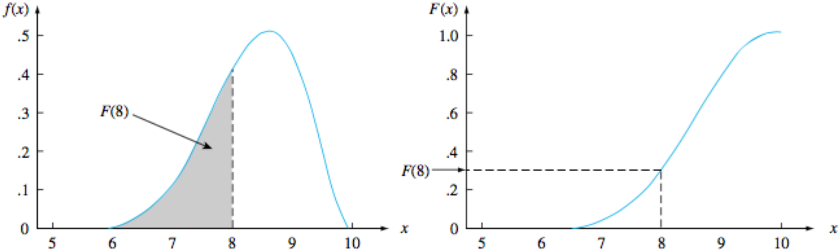


Figure 4.5 A pdf and associated cdf

Using CDF to compute probability

Let X be a continuous rv with pdf $f(x)$ and cdf $F(x)$. Then for any number a,

$$P(X > a) = 1 - F(a)$$

and for any two numbers a and b with $a < b$,

$$P(a \leq X \leq b) = F(b) - F(a)$$

Expected values

Expected value (discrete r.v.)

Let X be a discrete rv with set of possible values D and pmf $p(x)$. The **expected value** or **mean value** of X, denoted by $E(X)$ or μ_X, is

$$E(X) = \mu_X = \sum_{x \in D} x \cdot p(x)$$

This expected value will exist provided that $\sum_{x \in D} |x| \cdot p(x) < \infty$.

Expected value (discrete r.v.)

x	1	2	3	4	5	6	7
$p(x)$	.01	.03	.13	.25	.39	.17	.02

Expected value:

$$
\begin{aligned}
\mu &= 1 \cdot p(1) + 2 \cdot p(2) + \cdots + 7 \cdot p(7) \\
&= (1)(.01) + 2(.03) + \cdots + (7)(.02) \\
&= .01 + .06 + .39 + 1.00 + 1.95 + 1.02 + .14 = 4.57
\end{aligned}
$$

Expected value of a function (discrete r.v.)

If the rv X has a set of possible values D and pmf $p(x)$, then the expected value of any function $h(X)$, denoted by $E[h(X)]$ or $\mu_{h(X)}$, is computed by

$$E[h(X)] = \sum_{D} h(x) \cdot p(x)$$

assuming that $\sum_D |h(x)| \cdot p(x)$ is finite.

Expected value of a function (discrete r.v.)

Proposition:

$$E(aX + b) = a \cdot E(X) + b$$

Corollary:

Proof

$$E(aX + b) = \sum_D (ax + b) \cdot p(x) = a \sum_D x \cdot p(x) + b \sum_D p(x)$$
$$= aE(X) + b$$

1. For any constant a, $E(aX) = a \cdot E(X)$ [take $b = 0$ in (3.12)].
2. For any constant b, $E(X + b) = E(X) + b$ [take $a = 1$ in (3.12)].

Let X have pmf $p(x)$ and expected value μ. Then the **variance** of X, denoted by $V(X)$ or σ_X^2, or just σ^2, is

$$V(X) = \sum_D (x - \mu)^2 \cdot p(x) = E[(X - \mu)^2]$$

The **standard deviation** (SD) of X is

$$\sigma_X = \sqrt{\sigma_X^2}$$

Alternative formula:

$$V(X) = \sigma^2 = \left[\sum_D x^2 \cdot p(x)\right] - \mu^2 = E(X^2) - [E(X)]^2$$

Variance of a function

Rules of Variance:

$$V[h(X)] = \sigma^2_{h(X)} = \sum_D \{h(x) - E[h(X)]\}^2 \cdot p(x)$$

Property

$$V[h(X)] = \sigma^2_{h(X)} = \sum_D \{h(x) - E[h(X)]\}^2 \cdot p(x)$$

Expected value (continuous r.v.)

The **expected** or **mean value** of a continuous rv X with pdf $f(x)$ is

$$\mu_X = E(X) = \int_{-\infty}^{\infty} x \cdot f(x)\, dx$$

This expected value will exist provided that $\int_{-\infty}^{\infty} |x| f(x)\, dx < \infty$.

Expected value (continuous r.v.)

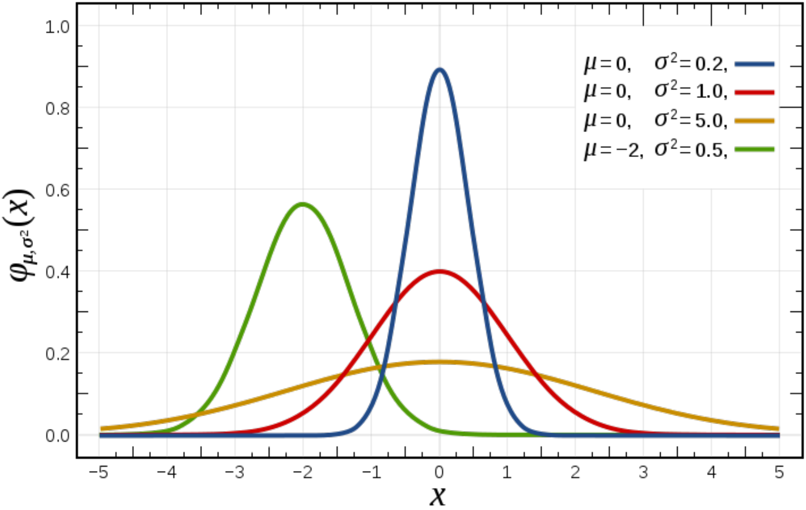
The **expected** or **mean value** of a continuous rv X with pdf $f(x)$ is

$$\mu_X = E(X) = \int_{-\infty}^{\infty} x \cdot f(x)\, dx$$

This expected value will exist provided that $\int_{-\infty}^{\infty} |x| f(x)\, dx < \infty$.

Normal distribution

$\mathcal{N}(\mu, \sigma^2)$



Basic properties

- $E(X) = \mu, \; Var(X) = \sigma^2$
- Density function

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

- $Z = \mathcal{N}(0, 1)$ is called the *standard normal distribution*

Standard normal distribution

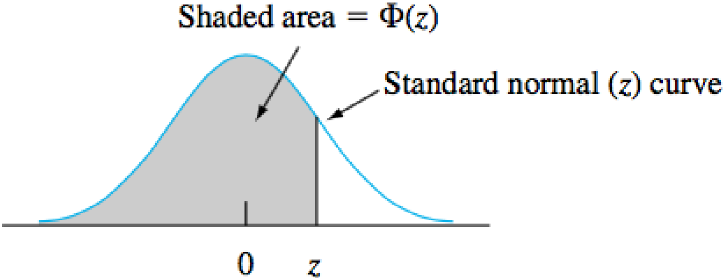
- $E(Z) = 0, \; Var(Z) = 1$
- Density function

$$f(z,0,1) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$$

- The cumulative distribution function of the standard normal distribution is:

$$\Phi(z) = P(Z \leq z) = \int_{-\infty}^{z} f(y,0,1) \; dy$$

$\Phi(z)$



$$\Phi(z) = P(Z \leq z) = \int_{-\infty}^{z} f(y, 0, 1)\; dy$$