Statistical Methods: Introduction, Applications, Histograms, Characteristics

Mike Renfro

November 4, 2004
Statistical Methods

Introduction

Application: How Often Will This Elevator Fail?

Definitions
Statistical Methods

Histograms and Probability Distribution Functions

Statistical Characteristics
- Measures of Central Tendency
- Measures of Dispersion
- Measures of Symmetry

Homework
Part I

Statistical Methods: Introduction and Applications
Introduction

- Engineering and scientific work often requires the collection of measurements or results from experiments.
- Most of the time, you can’t test every product. Sometimes because of the expense or time required to perform the test, sometimes because the test is destructive.
- But if the data collected is accurate and representative of all the tested and untested products, you can use your that data to draw conclusions about the population as a whole.
Simplified version: an elevator is supported by a wire rope that can withstand a stress of 30,000 psi. The load on the elevator causes a stress in the rope of 25,000 psi. Given this information alone, does the rope break?

No. As long as the stress from the elevator load is less than the rope strength, we’re ok. No lawsuits, no death or dismemberment, no need for safety devices, no problem.

Unfortunately, if you leave your analysis at this level, you’re very likely to have a big component failure at some point. Let’s try a less simplified example.
Application/Example: How Often Will This Elevator Fail?

- Statistical version: an elevator is supported by a wire rope that can withstand a stress of mean stress of 30,000 psi with a standard deviation of 1000 psi. The rope supplier isn’t perfect, and can’t provide ropes that have dead-on textbook values for the material’s yield and ultimate strengths 100% of the time.

- The load on the elevator is random, and causes a stress on the rope with a mean value of 25,000 psi and a standard deviation of 2000 psi.

- Without getting ahead of ourselves, let’s tentatively define the mean as the value that a measurement centers around, and that the measurements will be within $2\sigma$ of the mean around 95% of the time, where $\sigma$ is the standard deviation.

- So, given all this extra detail, are there situations where the rope can break?
We can definitely find situations where the elevator can fail.

Around 2.5% of the time, the stress on the rope will be higher than 29,000 psi.

Around 2.5% of the time, the rope will be 28,000 psi or weaker.

If you happen to subject an unusually weak rope to an unusually heavy load, it will break.

Qualitatively, if you plotted the probability distribution of the rope strength and the stress caused by the elevator load, wherever they overlapped would be where failures could occur.
Application/Example: How Often Will This Elevator Fail?

Example 6.1

- Distribution of Load
- Distribution of Strength
- Chance of Failure

Mike Renfro Statistical Methods: Introduction, Applications, Histograms, Characteristics
Definitions

- **Statistics.** A branch of mathematics dealing with the scientific method of collecting, analyzing, and displaying data. It also helps in developing models for describing and predicting physical phenomena.

- **Population.** The collection of all possible objects with a common measurable or observable feature or characteristic (strength, diameter, defects, etc.).

- **Sample.** A group of randomly-selected members of the population. All members of the sample are measured or tested in some repeatable way.

- **Sample size.** The number of members in the sample. Generally, larger sample sizes provide more accurate descriptions, and allow for more accurate predictions.
Definitions

- **Experiment.** The act of performing some thing where the exact outcome is not known. Examples: measuring the diameter of a shaft, tossing a coin, or examining a part for particular types of defects.

- **Event.** The outcome of an experiment. Some experiments have finite numbers of discrete outcomes (0 defects found, 1 defect found, 2 or more defects found). Others have an infinite number of real-valued outcomes (shaft diameter of 1.994 in, 1.989 in, 2.011 in, ⋯).

- **Random variable.** The measured or observed characteristic of a sample member. If the characteristic isn’t really random, it’s not a good candidate for statistical methods.
Definitions

- **Probability.** The chance of a particular event or range of events occurring. The probability of tossing a coin and it coming up heads is 50%. The probability of finding a shaft with a diameter less than 1.990 inches might be 3%, for example.
Part II

Statistical Methods: Histograms, Characteristics
Consider a random variable $X$ that take on $m$ discrete values.

If the experiment is repeated $n$ times, you'll find that the variable takes the values $x_1, x_2, x_3, \cdots, x_m$ a total of $n_1, n_2, n_3, \cdots, n_m$ times, respectively.

A bar chart with values of $x_i$ on the horizontal axis and $n_i$ on the vertical axis is a **histogram** or probability mass function.
Consider a random variable $X$ that be classified into $m$ discrete classes. Each class has a lower limit of $x_i - \frac{\Delta x}{2}$ and an upper limit of $x_i + \frac{\Delta x}{2}$.

If the experiment is repeated $n$ times, you’ll find that the variable falls into class $1, 2, 3, \cdots, m$ a total of $n_1, n_2, n_3, \cdots, n_m$ times, respectively.

A bar chart with values of $x_i$ on the horizontal axis and $\frac{n_i}{n}$ on the vertical axis is a probability density function.
Probability Density Function

Histogram (7 bins)

Yield Strength (kpsi) vs. % Occurrences

27 28 29 30 31 32 33
0
0.05
0.1
0.15
0.2
0.25
0.3
0.35
Mike Renfro
Statistical Methods: Introduction, Applications, Histograms, Characteristics
Cumulative Distribution Function

The cumulative distribution is found by adding up the bars on the probability density function from $x_1$ to $x_i$. 

![Cumulative Distribution Function](image_url)
Measures of Central Tendency

- **Mean.** Also known as the average, the mean of a set of $n$ data values $x_1, x_2, \cdots, x_n$ is calculated as

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

If the data are already separated into $m$ classes, as for a histogram, compute the mean as

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{m} x_{im} n_i$$

where $x_{im}$ is the midpoint of class $i$ and $n_i$ is the number of values of $x$ in class $i$. 
Measures of Central Tendency

- **Median.** The median is the value of $X$ where values above and below it occur with equal probability. If $n$ sorted data points are available, then $X_{\text{med}}$ can be determined as the middle value of $X$ if $n$ is odd, and the average of the two middle values of $X$ if $n$ is even.
Measures of Central Tendency

- **Mode.** The value of a random variable that is most likely to occur. If the data are separated into $m$ classes, the mode of $X$ can be determined as

$$X_{mod} = x_i \quad \text{with largest value of } n_i \ (i = 1, 2, \ldots, m)$$

- Modes are not as useful as means or medians in many cases. If the data $X$ is a continuous measurement rather than a discrete one, the choice of class boundaries can change the location of the mode.
Measures of Dispersion

- The **sample variance** of a random variable $X$ is denoted by $s^2$, and is a function of the deviations of the $x$ values from the mean $\bar{X}$:

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{X})^2 = \frac{1}{n-1} \left\{ n \sum_{i=1}^{n} x_i^2 - n\bar{X}^2 \right\}$$

- The **sample standard deviation** of the random variable is the positive square root of $s^2$:

$$s = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{X})^2}$$
Measures of Dispersion

- The **population variance** and **population standard deviation** of a random variable $X$ are used when the data represents the entire population rather than a smaller sample. They are denoted with a Greek lowercase sigma ($\sigma$) instead of lowercase “s”:

\[ \sigma^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{X})^2 \]

\[ \sigma = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{X})^2} \]
A skewed distribution is not symmetric, and many of the tests and procedures shown next time will not work properly on a skewed distribution. The basic skewness is calculated as

$$\text{Skewness} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{X})^3$$

and a skewness coefficient $\gamma$ is defined as

$$\gamma = \frac{\text{Skewness}}{s^3}$$
Work Problem 6.3. Experiment with histograms of this data in MATLAB. Assuming that you have your data in a vector variable named `beamLoads`, the following code will make a histogram with bins centered on `binCenters`:

```matlab
binCenters = 130:180;
[n, x] = hist(beamLoads, binCenters);
bar(x, n, 1);
```

Can you find a value for `binCenters` that has a bell-shaped curve, similar to one of the graphs in Figure 6.5?