Intro to statistics

Continued

Population variance

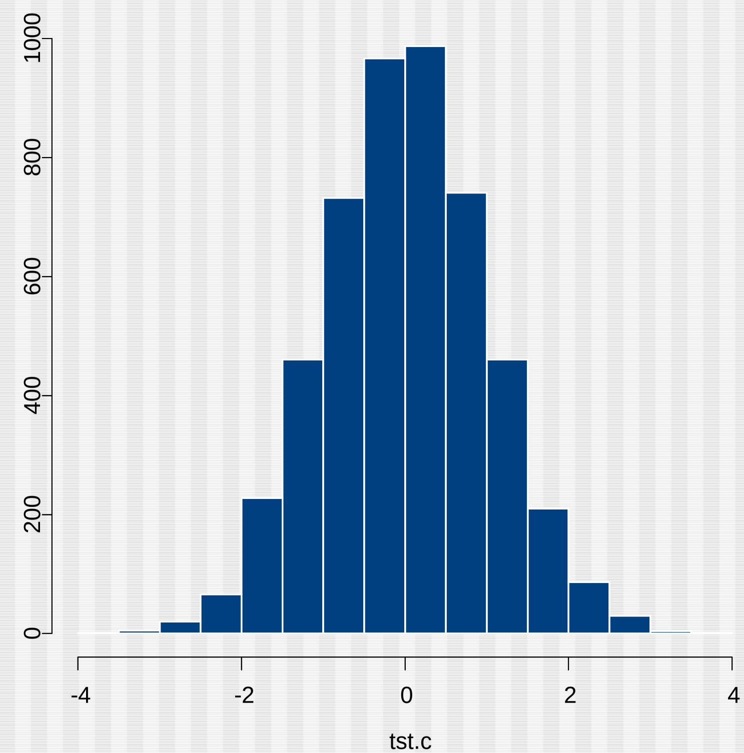
- Requires knowing μ (mean of population): $\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2$
- This if fine if we know about all the members of a population (so we can know the mean)
- What if we're calculating variance of a sample, and using it to estimate population variance?

Sample variance

- Use sample mean instead of population mean: $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$
- Calculated from samples, used as estimator of σ^2 (more detail later)
- Use "n-1" instead of "n" in denominator
- As n grows, the 2 calculations become closer

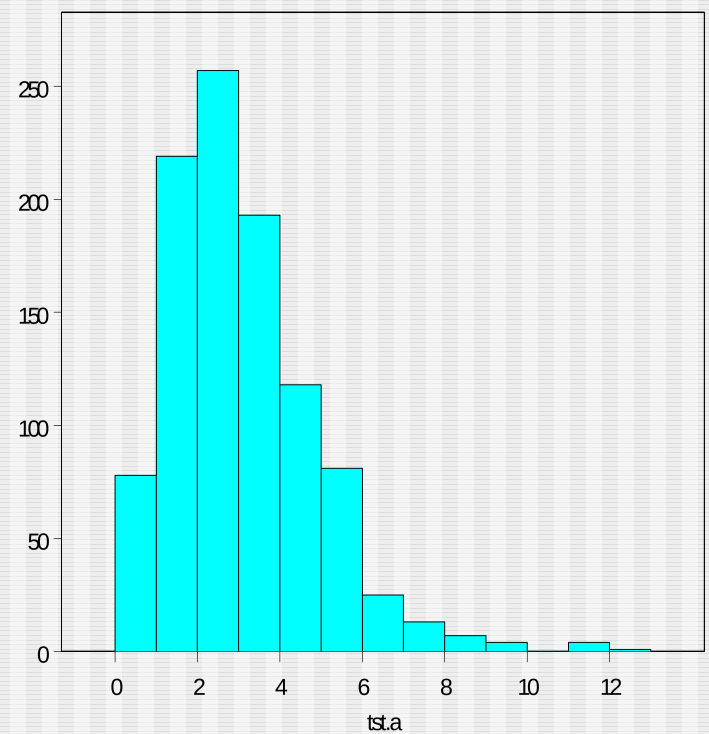
Symmetry

- Mean and median differ by < 1/100
- When mean = median, population is *symmetric*
- Think: one half is mirror image of other -> symmetric



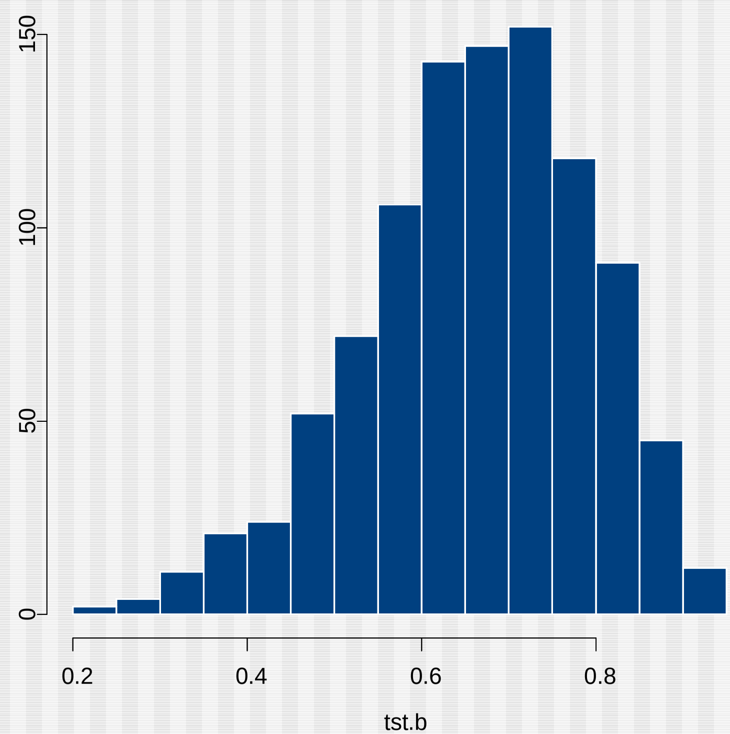
Right skew

- Mean is 3.07, median is 2.76
- Since mean > median, population has a *right skew*
- Think: tail extends to the right -> right skew

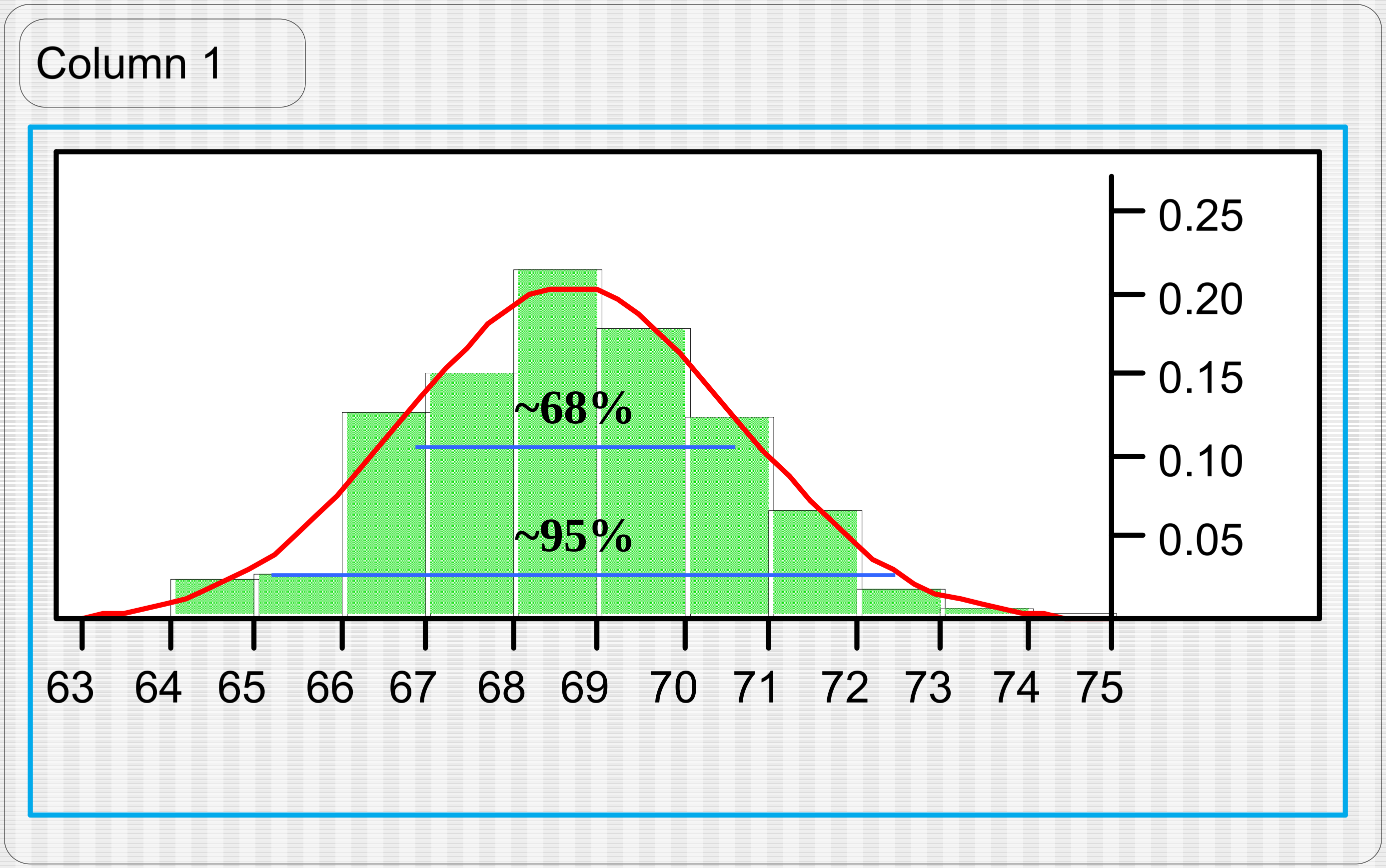


Left skew

- Mean is 0.6625, median is 0.673
- Since mean < median, population has a *left skew*
- Think: tail extends to the left -> left skew



Empirical rule



Experimental design

A brief intro

Basics

- Determine the question of interest
- Identify the response to be measured/observed
- Identify the treatment factor
- Define the treatment group and the control group
- Conduct study and compare groups

Treatment vs. control

- Observational study: Experimenter doesn’t assign subjects to treatment or control
- Experimental study: Experimenter does assign subjects to groups
- Groups should be as similar as possible (except for treatment factor)

Controlled experiments

- For best results, randomly choose subjects to be placed in treatment/control
- Avoids bias of experimenter in assigning groups
- Avoids self-selection bias

Other techniques/terms

- Placebo: “fake” treatment makes it impossible for subjects to know they are controls
- Double-blind study: designed so neither the subjects nor the experimenters who interact with them know which are controls

Observational study

- Experimenter cannot choose groups (practically/ethically)
- More difficult to establish similar treatment/control groups
- Can only establish association, not causation

Confounding variables

- Differences between groups that affect (are confounded with) the response
- Particularly problematic because confounding factors can be hard to spot

Ex: Ultrasounds and low birthweight

Several experiments on animals have shown that ultrasound exams can cause low birthweight.

Investigators ran an observational study to find out whether this is true for humans.

Ultrasounds and low birthweight (cont.)

- Researchers found some confounding variables and controlled for them
- Association was still found – babies exposed to ultrasound had lower birthweights on average
- Is this evidence that ultrasound causes low birthweight?

Ex: Coronary bypass surgery

In one early trial of coronary bypass surgery, a physician performed the surgery on a test group; 98% of whom survived at least 3 years.

Previous studies showed that 68% survived this long with conventional treatment.

Coronary bypass surgery (cont.)

- A newspaper commented on the physician’s results as “spectacular”.
- What, if anything, do you conclude?

Ex: U.S. murder rates

In 1985, 19,893 people were murdered, compared to 16,848 in 1970 (nearly 20% increase).

“These figures show that the U.S. became a more violent society over the period 1970-1985”.

True or false? Explain.

Background reading

- Examples were taken from *Statistics* by Freedman, Pisani, and Purves. (3rd ed., 1998)
- Chapters 1 and 2 contain more explanation and examples.

Probability

Basic ideas and concepts

What is it?

- Measure of belief in how likely something is to occur
- For unpredictable phenomenon, can say probability is relative frequency of event over the long haul

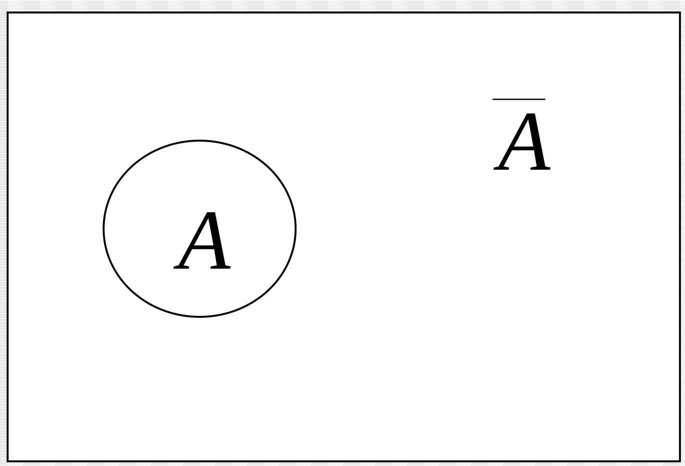
Short-term vs. long-term

- Would see some stable pattern if repeat experiment over and over -> random or stochastic events
- But at any point, we don’t know what result the experiment will yield next.

Probability vs. inference

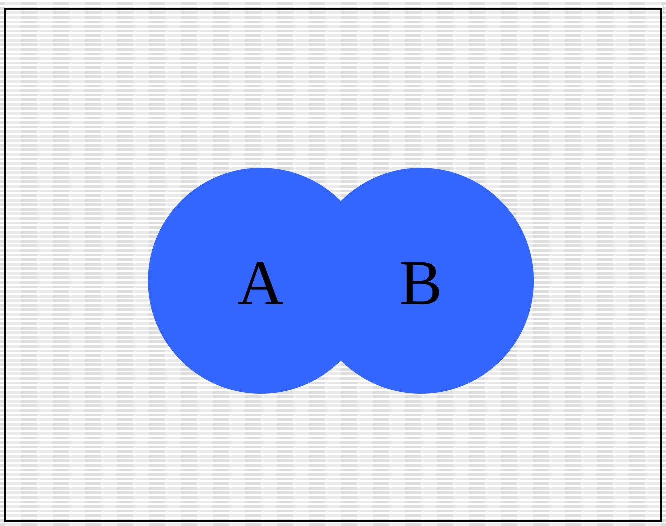
- If a coin is tossed 100 times, you would expect to see 50 heads
- However, there is some (non-zero) probability that it only comes up heads 25 times
- If there were 100 flips, and it only game up heads 25 times, you might *infer* that it wasn’t a fair coin.

Set notation



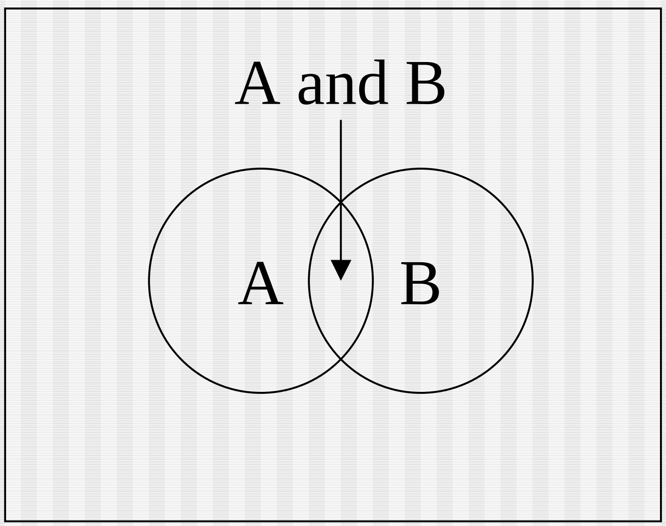
- Rectangle: set of all possible events
- Circle: subset of events producing the result A
- All events that don’t produce A are part of “A complement”

Union of events



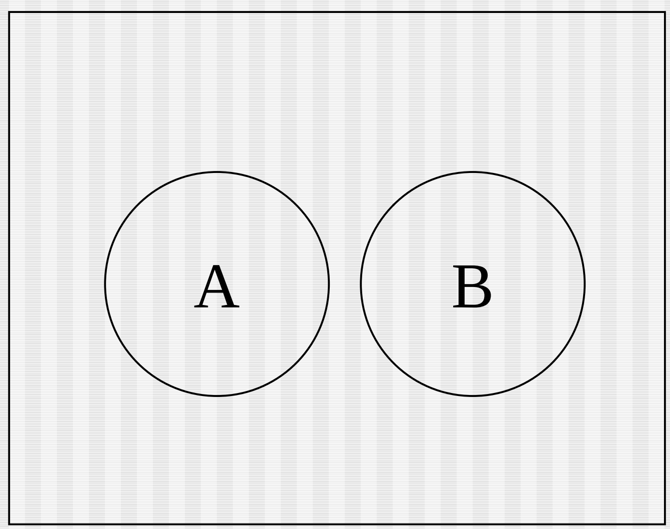
- Shaded area represents "union" of events A and B
- This represents set of events where A, or B, or both result
- Can be written "A or B" or "A U B"

Intersection of events



- Area where circles overlap represents "intersection" of events A and B
- Set of events where both A and B occur
- Can be written "A and B" or "A∩ B"

Disjoint events



- Since no overlap of circles, no events where both A and B occur
- A and B cannot happen at same time
- A and B also called “mutually exclusive” events