STAT 135, Concepts of Statistics

Helmut Pitters

Comparing two populations - matched samples

Department of Statistics
University of California, Berkeley

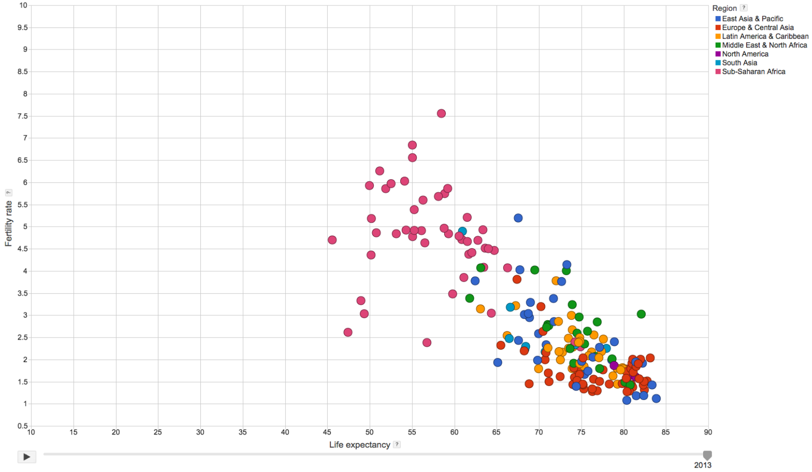
April 17, 2017

Review: Covariance and correlation.

Example: comparing production methods

Often in statistics not only interested in one random variable/population, but in relationships between different populations. – What if populations are not independent?

Relationship between two quantitative variables are usually displayed via scatterplots.



Scatterplot from Google Public Data Explorer.

Review: Covariance and correlation.

Recall from STAT 134: *Covariance* of two random variables X and Y is defined by

$$\mathrm{Cov}(X,Y) \coloneqq \mathbb{E}\left[(X-\mathbb{E}[X])(Y-\mathbb{E}[Y])\right].$$

The covariance can be interpreted as a measure for joint variability or degree of linear association.

If $\mathrm{Cov}(X,Y)=0$, we called X and Y *uncorrelated*. While independence of X, Y implies $\mathrm{Cov}(X,Y)=0$, the converse is not true.

Review: Covariance and correlation.

Recall from STAT 134 some useful formulas:

$$\begin{aligned}
\operatorname{Cov}(X, X) &= \operatorname{Var}(X) \\
\operatorname{Cov}(X, Y) &= \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] \\
\operatorname{Var}(X + Y) &= \operatorname{Var}(X) + \operatorname{Var}(Y) + 2\operatorname{Cov}(X, Y) \\
\operatorname{Cov}(X, Y) &= \operatorname{Cov}(Y, X) \qquad \text{(symmetry)} \\
\operatorname{Cov}(aX + b, Y) &= a\operatorname{Cov}(X, Y) \qquad \text{(multilinearity)}
\end{aligned}$$

Review: Covariance and correlation.

A drawback of the covariance is that it depends on the units in which X and Y are measured. We therefore agreed to first transform a random variable X to standard units, i.e. we center and rescale X

$$X^* := \frac{X - \mathbb{E}[X]}{\sqrt{\operatorname{Var}(X)}}$$

to standard units.[1] Accordingly, we defined the *correlation* of X and Y by

$$\begin{aligned}
\operatorname{Corr}(X, Y) := \operatorname{Cov}(X^*, Y^*) &= \frac{\operatorname{Cov}(X, Y)}{\sqrt{\operatorname{Var}(X)\operatorname{Var}(Y)}} \\
&= \frac{\mathbb{E}\left[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])\right]}{\sqrt{\operatorname{Var}(X)\operatorname{Var}(Y)}} = \mathbb{E}[X^* Y^*].
\end{aligned}$$

[1] In particular, $\mathbb{E}X^* = 0$, $\operatorname{Var}(X^*) = 1$.

Review: Covariance and correlation.

We found that for any two real random variables X and Y

$$-1 \leq \mathrm{Corr}(X, Y) \leq 1.$$

Moreover, $\mathrm{Corr}(X, Y) = 1$ or $= -1$ implies the existence of reals a, b such that

$$Y = aX + b.$$

Review: Covariance and correlation.

For a sample $(x_1, y_1), \dots, (x_n, y_n)$ of n paired observations the *sample covariance* is defined as

$$s_{xy} := \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}).$$

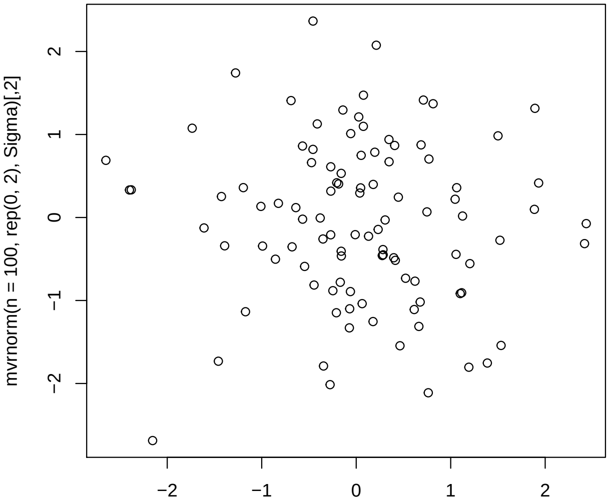
In complete analogy to the case of random variables, the *sample correlation coefficient* is defined as

$$r := \frac{s_{xy}}{s_x s_y},$$

where $s_x^2 := \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$, respectively s_y denotes the *sample variance* of the x-, respectively y-sample.

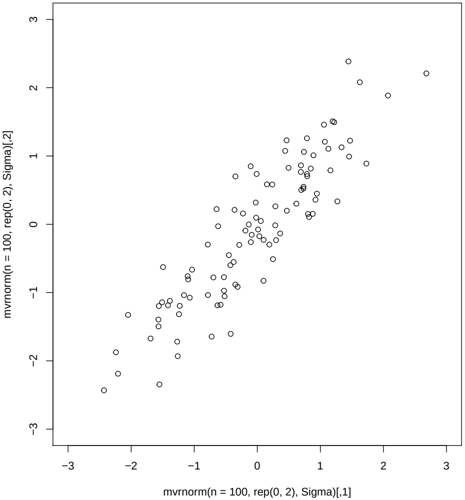
Review: Covariance and correlation.

100 samples from bivariate Normal distribution, $r = 0$.



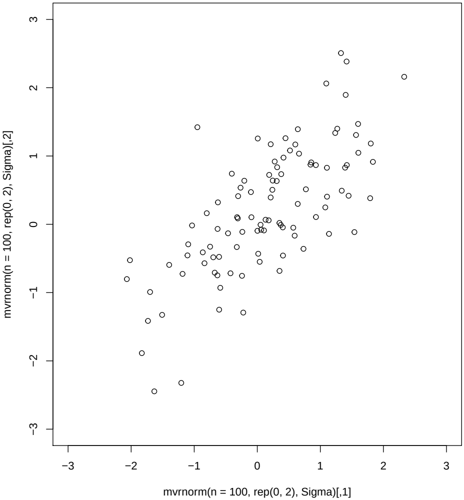
Review: Covariance and correlation.

100 samples from bivariate Normal distribution, $r = 0.9$.



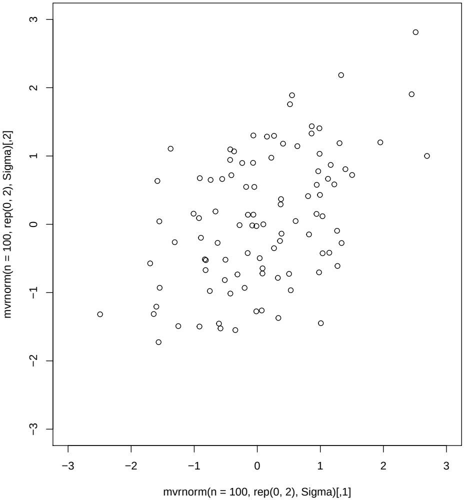
Review: Covariance and correlation.

100 samples from bivariate Normal distribution, $r = 0.7$.



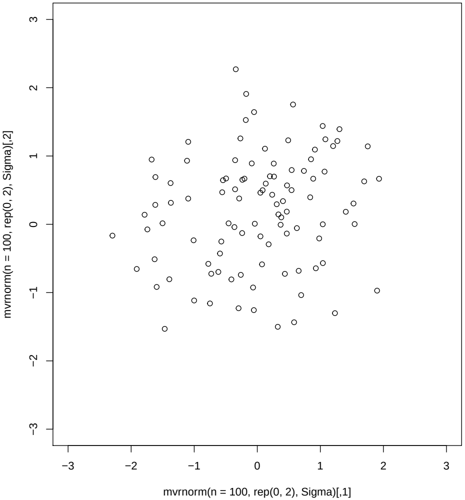
Review: Covariance and correlation.

100 samples from bivariate Normal distribution, $r = 0.5$.



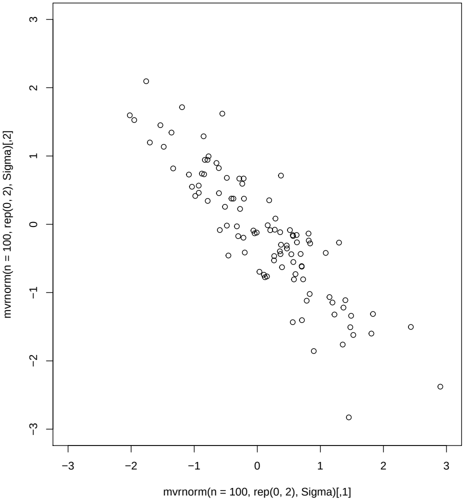
Review: Covariance and correlation.

100 samples from bivariate Normal distribution, $r = 0.2$.



Review: Covariance and correlation.

100 samples from bivariate Normal distribution, $r = -0.9$.



Comparing two populations. Matched samples

Back to comparing two populations. – Why matched pairs?

Example (Comparing production methods)

Want to compare two production methods. Each of $n = 6$ workers completes task once by method 1, and once by method 2.

$$\text{Completion times} \quad (t_i^1, t_i^2) \quad \text{for each worker } i \in \{1, \ldots, n\}$$

are recorded (in minutes).

method 1 (t_i^1)	method 2 (t_i^2)	difference ($t_i^1 - t_i^2$)
6.0	5.4	$.6$
5.0	5.2	$-.2$
7.0	6.5	$.5$
6.2	5.9	$.3$
6.0	6.0	$.0$
6.4	5.8	$.6$

Table: Completion times for method 1 and 2.

Comparing two populations. Matched samples

Example (Comparing production methods)

A meaningful procedure that compares method 1 and method 2 will be based on

$$\text{differences in completion times} \quad t_i^1 - t_i^2.$$

Completion times not only depend on production method, but also on worker. Want to eliminate the effect of worker speed on differences $t_i^1 - t_i^2$ (and thus reduce their variance).
Not matching completion times

$$t_1^1, t_2^1, t_3^1, \ldots, t_n^1, t_1^2, t_2^2, \ldots, t_n^2$$

corresponds to having $2n$ workers completing tasks, which introduces additional randomness (or noise) due to individual worker speed.

Comparing two populations. Matched samples

Setup and notation. Samples $(x_1, y_1), \dots, (x_n, y_n)$ assumed to be observations from n pairs

$$(X_1, Y_1), \dots, (X_n, Y_n) \quad \text{of i.i.d. random variables.}$$

Conceptually, comparing matched samples is easier than comparing non-paired samples from two populations. Why?
Can study

$$\text{differences} \qquad D_i := X_i - Y_i.$$

(If samples were not matched, which of the $n_1 n_2$ differences $X_i - Y_j$ should we consider?)

Comparing two populations. Matched samples

Population parameters

$$
\begin{aligned}
\mu_X &:= \mathbb{E}X_1, \qquad \mu_Y := \mathbb{E}Y_1 \\
\sigma_X^2 &:= \mathrm{Var}(X_1), \qquad \sigma_Y^2 := \mathrm{Var}(Y_1) \\
\sigma_{XY} &:= \mathrm{Cov}(X_i, Y_i) = \rho\sigma_X\sigma_Y,
\end{aligned}
$$

where $\rho := \mathrm{Corr}(X_i, Y_i)$. Consequently[2]

$$
\begin{aligned}
\mathbb{E}D_i &= \mathbb{E}[X_i - Y_i] = \mu_X - \mu_Y \\
\mathrm{Var}(D_i) &= \mathrm{Var}(X_i - Y_i) = \mathrm{Var}(X_i) + \mathrm{Var}(Y_i) - 2\,\mathrm{Cov}(X_i, Y_i) \\
&= \sigma_X^2 + \sigma_Y^2 - 2\rho\sigma_X\sigma_Y.
\end{aligned}
$$

[2] Generally, what can you say about noise in $X \pm Y$ based on σ_X, σ_Y, and ρ?

Comparing two populations. Matched samples

As before, interested in null hypothesis

$$H_0\colon \mu_X = \mu_Y \quad \text{or} \quad \mu_X - \mu_Y = 0$$

that populations (treatment and control) do not differ,
i.e. treatment has no effect, vs. $H_A\colon \mu_X \neq \mu_Y$.
As estimator for $\mu_X - \mu_Y$ (and test statistic) take

$$\bar{D}_n := \frac{1}{n}\sum_{i=1}^{n} D_i = \bar{X}_n - \bar{Y}_n$$

with

$$\mathbb{E}\bar{D}_n = \mu_X - \mu_Y \qquad \mathrm{Var}(\bar{D}_n) = \frac{1}{n}(\sigma_X^2 + \sigma_Y^2 - 2\rho\sigma_X\sigma_Y).$$

In principle we are done: as long as we can work out distribution of $\bar{D}_n$, we can

- work out confidence intervals
- do hypothesis tests, etc.

Comparing two populations. Matched samples

$$\mathbb{E}\bar{D}_n = \mu_X - \mu_Y \qquad \mathrm{Var}(\bar{D}_n) = \frac{1}{n}(\sigma_X^2 + \sigma_Y^2 - 2\rho\sigma_X\sigma_Y).$$

If samples were taken independently, without matching

$$\mathrm{Var}(\bar{X}_n - \bar{Y}_n) = \frac{1}{n}(\sigma_X^2 + \sigma_Y^2).$$

Thus matching is more effective if $\rho > 0$, i.e. if populations are positively correlated.

Normal distribution

Comparing two populations. Matched samples

Now assume $D_1, \ldots, D_n \sim \mathcal{N}(\mu_D, \sigma_D^2)$. Let

$$\bar{d}_n := \frac{1}{n}\sum_{i=1}^{n} d_i \qquad s_{\bar{D}_n} := \frac{1}{n-1}\sum_{i=1}^{n}(d_i - \bar{d}_n)^2$$

If σ_D^2 is known, use statistic

$$\frac{\bar{D}_n - \mu_D}{\sigma_D} \sim \mathcal{N}(0,1)$$

for inference.
Otherwise, statistic

$$t := \frac{\bar{D}_n - \mu_D}{s_{\bar{D}_n}/\sqrt{n}} \sim t_{n-1}$$

follows Student's t distribution with $n-1$ df and can be used for inference.

Comparing two populations. Matched samples

Example (Comparing production methods)

If worker i is particularly quick at completing the task by method 1, we expect that this is partly due to him being a fast worker. Therefore expect him to also be quick (in relation to other workers) at completing the task by method 2.
More formally: expect completion times (T_i^1, T_i^2) to be positively correlated. This is why its meaningful to match samples.

Example

Since $n = 6$ we have under null hypothesis $\mu_D = 0$

$$t = \frac{\bar{D}_n}{s_{\bar{D}_n}/\sqrt{n}} \sim t_5.$$

Since $t_5(0.025) = -2.57$, a level $\alpha = 0.05$ test of

$$H_0\colon \mu_D = 0 \qquad \text{vs.} \qquad H_A\colon \mu_D \neq 0$$

has decision rule

$$\text{reject } H_0 \text{ if } t \leq -2.57 \text{ or } t \geq 2.57.$$

Comparing two populations. Matched samples

Example

From data compute

$$\bar{d}_n = \frac{1}{n}\sum_{i=1}^{n}(t_i^1 - t_i^2) = 0.3$$

$$s_{\bar{D}_n} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(d_i - \bar{d}_n)^2} = 0.335$$

hence

$$t = \frac{\bar{d}_n}{s_{\bar{D}_n}/\sqrt{n}} = 2.19$$

that is the test does not reject H_0. The difference in completion times we see in the data can be explained by chance due to random sampling (at level $\alpha = 0.05$).

Comparing two populations. Matched samples

Nonparametric tests. We discuss ideas of nonparametric tests applied to two different settings of matched samples without going into details.

1. Data do not following normal distribution
   idea: rank absolute values $|D_1|, \ldots, |D_n|$
   null hypothesis: populations are identical
   under null, $\boxed{D_i = X_i - Y_i}$ is symmetric about 0, and

   $$\sum_{i=1}^{n} \text{sgn}(D_i)\text{rank}(|D_i|) \quad \text{should be close to } 0,$$

   where sgn(x) denotes the sign of x.
   [Wilcoxon signed rank test]
2. Data are not quantitative. E.g. customers indicate their preference for one of two products.

Comparing two populations. Matched samples

Let D be a (continuous) real r.v., symmetric about 0, i.e. $D =_d -D$.

$$\mathbb{P}\left\{\mathsf{sgn}(D) = 1\right\} = \mathbb{P}\left\{D \geq 0\right\} = \frac{1}{2} = \mathbb{P}\left\{D < 0\right\} = \mathbb{P}\left\{\mathsf{sgn}(D) = -1\right\},$$

that is, $\mathsf{sgn}(D)$ has Bernoulli $1/2$ distribution on $\{-1, 1\}$.
Moreover, for any $x \in \mathbb{R}$

$$\begin{aligned}
\mathbb{P}\left\{\mathsf{sgn}(D) = 1, |D| > x\right\} &= \mathbb{P}\left\{D > x\right\} \\
&= \frac{1}{2}(\mathbb{P}\left\{D > x\right\} + \mathbb{P}\left\{-D > x\right\}) \\
&= \frac{1}{2}\mathbb{P}\left\{|D| > x\right\} \\
&= \mathbb{P}\left\{\mathsf{sgn}(D) = 1\right\}\mathbb{P}\left\{|D| > x\right\},
\end{aligned}$$

showing (with similar calculation for $\mathsf{sgn}(D) = -1$) that $\mathsf{sgn}(D)$ and $|D|$ are independent.

Comparing two populations. Matched samples

Back to our setting:

$$D_1, \ldots, D_n \text{ are i.i.d., continuous, symmetric about } 0.$$

Hence

$$(\mathsf{sgn}(D_1), \mathsf{sgn}(D_2), \ldots, \mathsf{sgn}(D_n)) \quad \text{and} \quad (|D_1|, |D_2|, \ldots, |D_n|)$$

are both i.i.d. and independent of each other, and since rank$(|D_i|)$ only depends on $(|D_1|, \ldots, |D_n|)$,

$$(\mathsf{sgn}(D_1), \ldots, \mathsf{sgn}(D_n)) \text{ and } (\mathsf{rank}(|D_1|), \ldots, \mathsf{rank}(|D_n|))$$

are independent of each other.